Data assimilation for linear parabolic equations: minimax projection method

Sergiy Zhuk, Frank Jason, Isabelle Herlin, Robert Shorten

SIAM J. SCI. COMPUT.
Vol. 37, No. 3, pp. A1174–A1196

DATA ASSIMILATION FOR LINEAR PARABOLIC EQUATIONS: MINIMAX PROJECTION METHOD*

SERGIY ZHUK†, JASON FRANK‡, ISABELLE HERLIN§, AND ROBERT SHORTEN¶

Abstract. In this paper we propose a state estimation method for linear parabolic partial differential equations (PDE) that accounts for errors in the model, truncation, and observations. It is based on an extension of the Galerkin projection method. The extended method models projection coefficients, representing the state of the PDE in some basis, by means of a differential-algebraic equation (DAE). The original estimation problem for the PDE is then recast as a state estimation problem for the constructed DAE using a linear continuous minimax filter. We construct a numerical time integrator that preserves the monotonic decay of a nonstationary Lyapunov function along the solution. To conclude, we demonstrate the efficacy of the proposed method by applying it to the tracking of a discharged pollutant slick in a two-dimensional fluid.

Key words. PDEs, DAEs, state estimation, minimax, projection, structure preservation

AMS subject classifications. 35K20, 65L60, 93E10

DOI. 10.1137/13094709X

1. Introduction. Applications often give rise to new theoretical directions for numerical analysis. One example of such an application is the tracking of environmental slicks resulting from a discharge of a pollutant into the ocean. A simple mathematical model describing transport phenomena is the linear advection-diffusion equation:

$$\partial_t I + \boldsymbol{v} \cdot \nabla I - \varepsilon \Delta I = f\,, \tag{1.1}$$

where $I(\boldsymbol{x}, t)$ models the concentration of a pollutant at time t and location $\boldsymbol{x}$ within a given domain Ω, $\boldsymbol{v}(\boldsymbol{x}, t)$ is a divergence-free vector-field representing the fluid flow, and $f(\boldsymbol{x}, t)$ represents model error. In fact, (1.1) describes the diffusive linear transport of the initial pollutant concentration $I(\boldsymbol{x}, 0) = I_0(\boldsymbol{x})$ in the flow $\boldsymbol{v}$. Widespread interest in pollutant tracking problems stems from concerns of environmental agencies, energy producers, and governments worldwide, namely, one is interested in combining sensor readings obtained in real time (for instance, satellite images) in a sensible manner with the solution of (1.1) to generate reliable predictions of the pollutant's transport. Tracking pollutants based on sensor information is challenging because, on the one hand, the measurements are sparse and noisy, and, on the other hand, there is incomplete knowledge of the fluid flow $\boldsymbol{v}$. In fact, $\boldsymbol{v}$ is usually inferred from a flow identification procedure (see, for instance, [11]), and the resulting identification error f is only quantifiable as a bounded signal. This motivates the work of the present paper. Specifically, we are concerned with the following problem: estimate

*Submitted to the journal's Methods and Algorithms for Scientific Computing section December 2, 2013; accepted for publication (in revised form) January 28, 2015; published electronically May 5, 2015.

http://www.siam.org/journals/sisc/37-3/94709.html

†IBM Research—Ireland, Damastown, Dublin 15, Ireland (sergiy.zhuk@ie.ibm.com).

‡Mathematical Institute, Utrecht University, P.O. Box 80010, 3508 TA Utrecht, The Netherlands (j.e.frank@uu.nl).

§INRIA, Domaine de Voluceau BP 105 78153 Le Chesnay, France (isabelle.herlin@inria.fr).

¶IBM Research—Ireland, Damastown, Dublin 15, Ireland and School of Electrical Engineering, University College Dublin, Belfield, Dublin 4, Ireland (robshort@ie.ibm.com).

the state $I(\boldsymbol{x},t)$ of the infinite dimensional system (1.1) given incomplete and noisy observations y_k associated to $I(\boldsymbol{x},t)$ by

$$y_k(t) = \int_\Omega g_k(\boldsymbol{x},t) I(\boldsymbol{x},t) d\boldsymbol{x} + e_k(t)\,, \tag{1.2}$$

where g_k models a sensor and e_k represents the measurement error.

There are two principal strategies to construct a numerical state estimate for a partial differential equation (PDE). The first one is to consider the PDE (1.1) as a dynamical system with an infinite dimensional state space [6], then derive the estimator for the infinite dimensional system analytically, and finally reduce the resulting estimator to a finite dimensional system in order to construct the numerical estimate. The second strategy is to reduce a PDE to a finite dimensional system and then estimate the state of the reduced system. The first strategy applies either the semigroup theory given in [18] or the concept of Sobolev spaces [7]. One such class of infinite dimensional state estimators is based on optimal control theory for PDEs [15]. For example, linear optimal estimates for parabolic PDEs were derived in [4, 17]. We emphasize that to compute an optimal "off-line estimate" one needs to discretize the so-called Euler–Lagrange equations. This entails the solution of a two-point boundary value problem for (1.1) and its adjoint. On the other hand, obtaining an optimal "on-line estimate" in the form of a linear filter requires solving a nonlinear operator equation which is, in fact, an infinite dimensional counterpart of the matrix Riccati equation [19]. Another class of state estimators based on Lyapunov stability theory is represented by stable infinite dimensional observers with nonoptimal gains [3].

In this paper we adopt the second strategy, namely, we reduce the PDE (1.1) to a finite dimensional system and then derive an estimator for the reduced model. In order to do this systematically and to fully incorporate the reduction method into the state estimation procedure, we propose an extension of the classical Galerkin projection method. Recall that the Galerkin projection method is built upon the following requirement (see [12, p. 43]):

$$\partial_t I_N + \boldsymbol{v}\cdot\nabla I_N - \varepsilon\Delta I_N - f \perp \operatorname{span}\{\varphi_1 \dots \varphi_N\}\,, \tag{1.3}$$

where $I_N = \sum_{i=1}^N a_i(t)\varphi_i$ approximates I solving (1.1) and $\{\varphi_1 \dots \varphi_N\}$ denotes a finite basis for the projection space. Condition (1.3) yields the following reduced system for the vector of approximated projection coefficients $\boldsymbol{a} = (a_1 \dots a_N)'$:

$$\langle \partial_t I_N + \boldsymbol{v}\cdot\nabla I_N - \varepsilon\Delta I_N - f, \varphi_n\rangle = 0\,, \quad n = 1,\dots,N\,. \tag{1.4}$$

If the exact solution I of (1.1) were known, we could project it onto span $\{\varphi_k\}_{k=1}^N$ to obtain the vector of the exact projection coefficients $\boldsymbol{a}_N^{true}$. However, we stress that $\boldsymbol{a}_N^{true}$ satisfies (1.4) in a very special case, namely, if $\operatorname{span}\{\varphi_1 \dots \varphi_N\}$ is invariant with respect to the differential operator $\mathcal{A} = \boldsymbol{v}\cdot\nabla - \varepsilon\Delta$. In the general case, the solution of ODE (1.4) deviates from $\boldsymbol{a}_N^{true}$ for any finite N as the system (1.4) is not closed, namely, it does not retain all the information which is necessary to describe the evolution of $\boldsymbol{a}_N^{true}$. Although the solution of (1.4) converges to $\boldsymbol{a}_N^{true}$ in the limit $N \to \infty$ provided the basis functions are consistent and stable (see [12, p. 251] for details), the limiting case $N \to \infty$ is less interesting for our purposes, as in operational practice one usually fixes N before deriving the state estimator.

In this paper we wish to explicitly account for the truncation error induced by (1.4), combining knowledge of the regularity properties of the parabolic operator $\mathcal{A}$

and information available from indirect observations of $\boldsymbol{a}_N^{true}$ to control the error of an approximate solution such that it resides within a known ellipsoid centered around $\boldsymbol{a}_N^{true}$. Thus, for a finite N, it is necessary for us to have a closed system describing the evolution of $\boldsymbol{a}_N^{true}$, in contrast to the Galerkin projection methods discussed in [2, 12, 16, 20], where the authors study the limiting behavior only of the approximation error for the nonclosed system (1.4). Consequently, we propose the following extension of the Galerkin method: we close (1.4) by including an unknown input vector field $\boldsymbol{e}^m$ that models the impact of the truncated coefficients $\{a_{N+1}, a_{N+2}, \dots\}$ on $\boldsymbol{a} = (a_1, \dots, a_N)'$. We allow $\boldsymbol{e}^m$ to take on values within an a priori determined bounding set constructed using parabolic regularity theory [7]. Further, we introduce an additional algebraic constraint that filters out inadmissible $\boldsymbol{e}^m$. Assembling the resulting differential and algebraic equations together, we arrive at a reduced model for $\boldsymbol{a}_N^{true}$ in the form of a differential algebraic equation (DAE). Formal derivation of the described reduction of PDE (1.1) to the DAE is given in Proposition 3.1. The formulation is contrasted with the classical Galerkin method in section 3.2.

Although the literature on projection methods and related control techniques is very rich, to the best of our knowledge the proposed extension of the Galerkin method is new and has not been discussed in the literature yet. A preliminary version of the minimax projection method appeared in [27]. With respect to [27], the main difference is that our state equation is time-dependent and has unknown parameters, and the time-discretization process has been incorporated into the state estimation procedure.

In our description, the model error f and input $\boldsymbol{e}^m$ are unknown deterministic functions that are bounded. The latter is natural as the input $\boldsymbol{e}^m$ represents the truncation error which can be expanded in Fourier series. On the other hand, the observation error e_k is assumed to be a realization of a random process which is the most commonly used uncertainty model for such errors. In practice, second moments of e_k are usually given with some error. Thus, the state estimator should be robust with respect to fluctuations of the covariance matrix. We propose to address this issue by assuming that the covariance function of e_k is unknown but bounded and belongs to a given ellipsoid. As a result, we need to deal with unknown bounded deterministic parameters in the state equation and random noise in the observation equation, which has an unknown but bounded covariance function.

Within the context just sketched, it is natural to analyze the worst-case realization of all the unknown parameters to derive the state estimate for $\boldsymbol{a}_N^{true}$. This may be realized using the minimax state estimation approach [5, 14, 17], which was extended to DAEs in [22, 23, 24, 25]. To construct the minimax state estimate for a DAE, we apply a generalized Kalman duality principle [26] that converts the state estimation problem into a dual control problem with quadratic cost. This cost is, in fact, the worst-case estimation error and so, by minimizing it, we get the state estimate with the minimal worst-case error. The dual control problem is derived in Proposition 3.6 and provides a key component of the estimation procedure as it accumulates all the information about the deterministic projection error $\boldsymbol{e}^m$, model error f, and random measurement error e_k, which are present in the original DAE. As a result, the solution of the dual control problem has minimal worst-case error and is, therefore, robust to all the aforementioned sources of errors. This solution provides a starting point for the design of a numerical minimax estimate. Namely, we represent the minimax estimate $\hat{\boldsymbol{a}}_N$ for $\boldsymbol{a}_N^{true}$ in the form of a linear filter. We also build an ellipsoid which is centered around $\hat{\boldsymbol{a}}_N$ and contains[1] $\boldsymbol{a}_N^{true}$. This ellipsoid describes how the DAE

[1] In fact, the latter inclusion holds only on average, as the measurements contain a random error.

propagates all admissible errors enclosed in the a priori determined bounding set constructed in Proposition 3.1. The ellipsoid is parametrized by a symmetric positive definite matrix-valued function $t \mapsto K_N(t)$, which is obtained by solving a differential Riccati equation (DRE). Further, the largest eigenvalue of $K_N(t)$ defines the worst-case estimation error. All these important details are summarized in Corollary 3.7.

To conclude the introduction, we turn our attention to the numerical method. As mentioned above, the key ingredients of the minimax estimation are the linear filter $\hat{\boldsymbol{a}}_N$ and matrix K_N defined in Corollary 3.7. It is worth noting that the DRE is well understood from the theoretical point of view: for example, it is known (see [19] for the details) that $K_N(t) = V(t)U^{-1}(t)$, where V and U solve an associated linear Hamiltonian ODE. The exact representation of this ODE is given in Corollary 3.7. We make use of this fact to approximate V and U using a generic s-stage symplectic Runge–Kutta (RK) method of order p. The corresponding numerical method is derived in Proposition 4.1. Symplectic RK methods preserve quadratic invariants [10], an important point as the proper choice of discretization method ensures that the structure of the estimation error is preserved for the discrete filter $\hat{\boldsymbol{a}}_N^n$. Namely, the estimation error admits a nonstationary Lyapunov function which is preserved by the proposed discretization, as is discussed in Remark 3. The importance of this is that the simulation results are trustworthy and represent indeed what has been predicted by the theory for the continuons case. To the best of our knowledge, this is the first result of this kind in the framework of minimax state estimation. We note that symplectic RK methods were applied in [8] to obtain structure preserving discretization of Möbius integrators for DREs arising in the context of control problems for ODEs. In this paper we generalize this result to DAEs. To illustrate our approach, we represent the discrete filter $\hat{\boldsymbol{a}}_N^n$ by means of the implicit midpoint rule and apply it to the tracking problem, assuming that a slick of a discharged pollutant moves in the flow generated by the two-dimensional (2D) incompressible Euler equation.

This paper is organized as follows. Subsection 1.1 gives the notation used in the paper. Section 2 presents the formal problem statement. Our main results are given in section 3: subsections 3.1 and 3.2 contain the minimax projection method and comparisons to the classical Galerkin approach. Subsection 3.3 derives the minimax estimate for the projection coefficients and provides the worst-case estimation error. Section 4 introduces the structure-preserving discretization for the minimax estimate. Finally, section 5 presents the case study, and conclusions are given in section 6.

1.1. Notation. $\mathbb{N}$ denotes the set of natural numbers $\{1, 2, \dots\}$; $\mathbb{R}^n$ denotes n-dimensional Euclidean space; $\boldsymbol{x}\cdot\boldsymbol{w}$ denotes the canonical inner product for $\boldsymbol{x}, \boldsymbol{w} \in \mathbb{R}^n$, $\|\boldsymbol{x}\|^2_{\mathbb{R}^n} := \boldsymbol{x}\cdot\boldsymbol{x}$ and, more generally, $\langle f, g\rangle_H$ denotes the canonical inner product in a Hilbert space H for $f, g \in H$ and $\|f\|^2_H := \langle f, f\rangle$; $L^2(0,T,H) := \{f : f(t) \in H$ and $\int_0^T \|f(t)\|^2_H\, dt < +\infty\}$; $L^\infty(0,T,H) := \{f : \operatorname{esssup}_{0\le t\le T} \|f(t)\|_H < +\infty\}$; Ω is an open subset of $\mathbb{R}^n$ with boundary $\partial\Omega$; $\Omega_T := \Omega \times (0,T)$; $\|f\|^2_{L^2(\Omega)} = \int_\Omega f^2(\boldsymbol{x})d\boldsymbol{x}$, where $L^2(\Omega)$ denotes the space of all measurable f such that $\|f\|^2_{L^2(\Omega)} < +\infty$.

$L^\infty(\Omega)$ is a space of measurable functions bounded almost everywhere in Ω; $C(\overline{\Omega})$ is a space of continuous functions over the closure $\overline{\Omega}$ of Ω; $C_c^\infty(\Omega)$ is a space of all infinitely differentiable functions with compact support in Ω; $\partial_{x_i} f$ denotes the weak derivative of f; ∇f denotes the spatial gradient of f; $\Delta f := \sum_{i=1}^n \partial^2_{x_i^2} f$; $H^1(\Omega) := \{f \in L^2(\Omega) : \nabla f \in L^2(\Omega)\}$; $\|f\|^2_{H^1(\Omega)} = \|f\|^2_{L^2(\Omega)} + \|\nabla f\|^2_{L^2(\Omega)}$; $H^1_0(\Omega)$ is a closure of $C_c^\infty(\Omega)$ with respect to the norm of $H^1(\Omega)$; $\mathcal{I}$ stands for an identity operator or matrix; $J_{\boldsymbol{v}}(\boldsymbol{x})$ denotes Jacobian matrix for the vector-field $\boldsymbol{v}$; $\|A\|_2^2 := \sum_{i,j=1}^{m,n} a_{i,j}^2$ is

the Frobenius norm for the matrix $A \in \mathbb{R}^{m\times n}$; the prime $'$ denotes the operation of taking the adjoint: $\mathcal{A}'$ denotes the adjoint operator and A' denotes the transposed matrix; $\|A\|$ is the largest singular value of A; $A^{\frac{1}{2}}$ denotes the square root of a symmetric semidefinite matrix A; $\delta_{i,j} = 1$ if $i = j$ and 0 otherwise. Finally, $\mathbb{E}$ denotes the expectation of a random variable in its associated probability measure.

2. Problem statement. In this section we formalize the program laid out in section 1 to specify a problem statement. Assume $\varepsilon > 0$ and that $I(\cdot,t) \in H_0^1(\Omega)$ satisfies for almost all $t \in (0,T)$ the following equation:

$$\partial_t I + \boldsymbol{v}\cdot\nabla I - \varepsilon\Delta I = f\,,\quad I(\boldsymbol{x},0) = I_0(\boldsymbol{x})\,,\quad I(\boldsymbol{x},t) = 0\,,\boldsymbol{x}\in\partial\Omega\,, \tag{2.1}$$

where $\boldsymbol{x} \in \Omega \subset \mathbb{R}^n$, $n \geq 2$, Ω is an open bounded convex set and $\boldsymbol{v}(\boldsymbol{x},t) = (M_1(\boldsymbol{x},t)\dots M_n(\boldsymbol{x},t))'$ with $M_i \in L^\infty(0,T,H_0^1(\Omega))$ for all $i = 1,\dots,n$.

Suppose also that the deterministic model error $f \in L^2(0,T,L^2(\Omega))$ and initial condition $I_0 \in H^2(\Omega)\cap H_0^1(\Omega)$ satisfy the following inequality:

$$\int_\Omega Q_0(\boldsymbol{x})\nabla I_0(\boldsymbol{x})\cdot\nabla I_0(\boldsymbol{x})d\boldsymbol{x} + \int_{\Omega_T} Q(\boldsymbol{x},t)f^2(\boldsymbol{x},t)d\boldsymbol{x}\,dt \leq 1\,, \tag{2.2}$$

where $Q_0(\boldsymbol{x})$ is a symmetric matrix such that $Q_0 \in C(\overline{\Omega})$ and $\underline{q}_0\|\boldsymbol{\xi}\|^2_{\mathbb{R}^n} \leq Q_0(\boldsymbol{x})\boldsymbol{\xi}\cdot\boldsymbol{\xi} \leq \overline{q}_0\|\boldsymbol{\xi}\|^2_{\mathbb{R}^n}$ for all $\boldsymbol{x}\in\Omega$, $\boldsymbol{\xi}\in\mathbb{R}^n$ and given $0 < \underline{q}_0 \leq \overline{q}_0 < +\infty$, and $Q \in C(0,T,C(\overline{\Omega}))$ is a weighting function such that $0 < \underline{q}(t) \leq Q(\boldsymbol{x},t) \leq \overline{q}(t) < +\infty$ for the given $\underline{q},\overline{q}$. We note that Q_0 and Q may be considered as design parameters which quantify our level of confidence in I_0 and f, namely, Q_0 may specify "zones" of Ω where our knowledge of the initial condition is more precise or less so, and Q defines zones of Ω where (2.1) holds almost exactly or only up to a significant error and these zones may vary over time.

We assume that a vector $\boldsymbol{y}(t) = (y_1(t)\dots y_M(t))'$ is observed in the form

$$y_k(t) = \int_\Omega g_k(\boldsymbol{x},t)I(\boldsymbol{x},t)d\boldsymbol{x} + e_k(t)\,,k = 1,\dots,M\,, \tag{2.3}$$

where $g_k \in L^2(0,T,L^2(\Omega))$ is a spatial averaging kernel that models the effect of a measurement instrument, and $\boldsymbol{e} = (e_1(t)\dots e_M(t))'$ is a realization of a random process with zero mean and unknown but bounded covariance function $\operatorname{cov}(t,s) := \mathbb{E}\boldsymbol{e}(t)\boldsymbol{e}'(s)$, that is,

$$\int_0^T \operatorname{trace}(R(t)\operatorname{cov}(t,t))\,dt \leq 1\,, \tag{2.4}$$

where $t \mapsto R(t)$ is a symmetric positive definite continuous matrix-valued function with bounded inverse. In other words, the covariance function of $\boldsymbol{e}$ belongs to an ellipsoid in the space of real symmetric positive definite matrices where the trace is taken as the inner product. We note that, in practice, second moments of e_k are usually given with some error, and this fact is reflected by assumption (2.4): indeed, the inequality (2.4) represents a constraint on the weighted second moments of $\boldsymbol{e}$ as $\int_0^T \operatorname{trace}(R\operatorname{cov}(t,t))\,dt = \mathbb{E}\int_0^T R\boldsymbol{e}\cdot\boldsymbol{e}\,dt$.

Now, assuming that functions $\{\varphi_k\}_{k\in\mathbb{N}}$ form an orthonormal basis in $L^2(\Omega)$, we expand the solution I into the following series:

$$I(\boldsymbol{x},t) = \sum_{i\in\mathbb{N}} a_i(t)\varphi_i(\boldsymbol{x})\,,\quad a_i(t) := \langle I(\cdot,t),\varphi_i\rangle_{L^2(\Omega)}\,. \tag{2.5}$$

Note that I is completely defined by the coefficients $\{a_i(t)\}_{i\in\mathbb{N}}$ and $I^N = \mathcal{P}_N^\dagger \mathcal{P}_N I$ represents the most natural approximation for I in the given basis $\{\varphi_k\}_{k=1}^N$, provided that

$$\mathcal{P}_N I(\cdot,t) = \boldsymbol{a}(t) = (a_1(t)\dots a_N(t))' \text{ and } \mathcal{P}_N^\dagger \boldsymbol{a}(t) = \sum_{i=1}^N a_i(t)\varphi_i\,. \tag{2.6}$$

The formal problem statement for our state estimation algorithm is the following:

(i) Construct matrices A_N, H_N, C_N and a bounding set $\mathscr{E}_N$ such that the vector of exact projection coefficients $\boldsymbol{a}_N^{true} := \mathcal{P}_N I$ satisfy for some $(I_0, f, \boldsymbol{e}^m, \boldsymbol{e}^o, \boldsymbol{w}) \in \mathscr{E}_N$ the following DAE:

$$\begin{aligned} \frac{d\boldsymbol{a}}{dt} &= -A_N\boldsymbol{a} + \boldsymbol{e}^m + \mathcal{P}_N f\,,\\ 0 &= H_N\boldsymbol{a} + \boldsymbol{e}^o\,, \boldsymbol{a}(0) = \mathcal{P}_N I_0\,,\\ \boldsymbol{y}(t) &= C_N(t)\boldsymbol{a} + \boldsymbol{w}(t) + \boldsymbol{e}(t)\,, \end{aligned} \tag{2.7}$$

where $(\boldsymbol{e}^m, \boldsymbol{e}^o, \boldsymbol{w})$ stand for the projection error. See section 3.1.

(ii) Design a linear minimax estimate $\hat{\boldsymbol{a}}_N$ for the state of (2.7), that is, a vector-valued function $t \mapsto \hat{\boldsymbol{a}}_N(t)$ such that $\boldsymbol{\ell}\cdot\hat{\boldsymbol{a}}_N(t) = \hat{\boldsymbol{u}}(\boldsymbol{y}) := \int_0^t \hat{\boldsymbol{u}}\cdot\boldsymbol{y}dt$ and

$$\mathbb{E}(\boldsymbol{\ell}\cdot\boldsymbol{a}_N^{true}(t) - \hat{\boldsymbol{u}}(\boldsymbol{y}))^2 \le \sigma(\hat{\boldsymbol{u}}, t, \boldsymbol{\ell}) \le \sigma(\boldsymbol{u}, t, \boldsymbol{\ell}) \quad \forall \boldsymbol{u} \in L^2(0,t)\,, \boldsymbol{\ell} \in \mathbb{R}^N\,, \tag{2.8}$$

where $\sigma(\boldsymbol{u}, t, \boldsymbol{\ell}) := \sup_{(I_0, f, \boldsymbol{e}^m, \boldsymbol{e}^o, \boldsymbol{w}) \in \mathscr{E}_N, \boldsymbol{e}} \mathbb{E}(\boldsymbol{\ell}\cdot\boldsymbol{a}(t) - \boldsymbol{u}(\boldsymbol{y}))^2$ is the estimation error, corresponding to the worst-case realizations of parameters $(I_0, f, \boldsymbol{e}^m, \boldsymbol{e}^o, \boldsymbol{w}) \in \mathscr{E}_N$ and observation error $\boldsymbol{e}$ satisfying (2.4). See section 3.3.

(iii) Introduce a discrete-time minimax estimate $n \to \hat{\boldsymbol{a}}_N^n$ such that (2.8) holds for $\hat{\boldsymbol{a}}_N^n$ and discretized error functional σ. See section 4.

In fact, $\hat{\boldsymbol{a}}_N$ represents a robust estimate of $\boldsymbol{a}_N^{true}$ with minimal worst-case estimation error σ and its discrete analogue $\hat{\boldsymbol{a}}_N^n$ has a minimal worst-case error in discrete time.

3. Minimax estimate for the projection coefficients.

3.1. Minimax projection method. We begin with specifying basis functions used in (2.6) to define the projection operator $\mathcal{P}_N$. Recall that the Laplacian operator $-\Delta$ possesses an orthonormal set of eigenfunctions $\{\varphi_k\}_{k\in\mathbb{N}}$ in $L^2(\Omega)$:

$$-\Delta\varphi_k = \lambda_k\varphi_k\,, \quad \varphi_k \in C^\infty(\Omega)\cap H_0^1(\Omega)\,, \quad \varphi_k = 0 \text{ on } \partial\Omega\,,$$

where $0 < \lambda_1 \le \lambda_2 \le \cdots$ and $\lim_{k\to\infty}\lambda_k = +\infty$ (see [7, p. 355]).

Define the differential operator $\mathcal{A}\varphi = \boldsymbol{v}\cdot\nabla\varphi - \varepsilon\Delta\varphi$ associated with (2.1) and its projection $A_N := \mathcal{P}_N\mathcal{A}\mathcal{P}_N^\dagger$. Note that $\mathcal{P}_N\mathcal{P}_N^\dagger = \mathcal{I}$ as the $\{\varphi_k\}_{k\in\mathbb{N}}$ are orthogonal. Let $C_N(t) := \{\langle g_k(\cdot,t), \varphi_s\rangle_{L^2(\Omega)}\}_{k,s=1}^{M,N}$ correspond to the projection of the observation operator. We also introduce $S_N = \{\langle\mathcal{A}\varphi_i, \mathcal{A}\varphi_j\rangle\}_{i,j=1}^N$ and set[2] $H_N := (S_N - A_N'A_N)^{\frac{1}{2}}$.

Next, we formulate the main theoretical result of this article. It states that the projection coefficients of the exact solution $I(\boldsymbol{x}, t)$ solve the finite dimensional DAE (2.7) depending on error terms that can be bounded within a certain ellipsoid $\mathscr{E}_N$.

PROPOSITION 3.1. *Assume that I solves* (2.1) *for some I_0 and f satisfying* (2.2) *and $\boldsymbol{y}(t)$ is associated to I through* (1.2). *Then there exist $\boldsymbol{e}^m \in \mathbb{R}^N$, $\boldsymbol{e}^o \in \mathbb{R}^N$, and*

[2]The square root of $S_N - A_N'A_N$ is well defined as follows from (3.7).

$\boldsymbol{w} \in \mathbb{R}^M$ such that the vector of the exact projection coefficients $\boldsymbol{a}_N^{true} = \mathcal{P}_N I \in \mathbb{R}^N$ solves the DAE

$$\begin{aligned} \frac{d\boldsymbol{a}}{dt} &= -A_N \boldsymbol{a} + \boldsymbol{e}^m + \mathcal{P}_N f\,, \\ 0 &= H_N \boldsymbol{a} + \boldsymbol{e}^o\,, \boldsymbol{a}(0) = \mathcal{P}_N I_0\,, \\ \boldsymbol{y}(t) &= C_N(t)\boldsymbol{a} + \boldsymbol{w}(t) + \boldsymbol{e}(t)\,, \end{aligned} \tag{3.1}$$

where $(I_0, f, \boldsymbol{e}^m, \boldsymbol{e}^o, \boldsymbol{w})$ belong to the ellipsoid $\mathscr{E}_N$:

(3.2)

$$\begin{aligned} \mathscr{E}_N := \Big\{ (I_0, f, \boldsymbol{e}^m, \boldsymbol{e}^o, \boldsymbol{w}) : \int_\Omega Q_0(\boldsymbol{x}) \nabla I_0(\boldsymbol{x}) \cdot \nabla I_0(\boldsymbol{x}) d\boldsymbol{x} + \int_{\Omega_T} Q(\boldsymbol{x}, t) f^2(\boldsymbol{x}, t) d\boldsymbol{x}\, dt \\ + \lambda_{N+1}^{\frac{1}{2}} \int_0^T \|S^{\frac{1}{2}} \boldsymbol{e}^m\|^2_{\mathbb{R}^N} + \|S^{\frac{1}{2}} \boldsymbol{e}^o\|^2_{\mathbb{R}^N} + \lambda_{N+1}^{\frac{1}{2}} \|V^{\frac{1}{2}} \boldsymbol{w}\|^2_{\mathbb{R}^M}\, dt \le \mu_N \Big\} \end{aligned}$$

for certain positive constants S, V and $\mu_N := 1 + \lambda_{N+1}^{-\frac{1}{2}} + \lambda_{N+1}^{-1}$, S and V.

We postpone the proof of this proposition until the end of this section.

Remark 1. On the one hand, the DAE (3.1) does not seem to be useful from the computational standpoint as $\boldsymbol{e}^m, \boldsymbol{e}^o$ are linear functions of $\boldsymbol{a}_N^{true}$ and the latter is unknown. Therefore, in practice, $\boldsymbol{e}^m$ and the second equation in (3.1) are usually dropped and only the first equation of (3.1) is used in numerical computations to approximate $\boldsymbol{a}_N^{true}$. However, if we change our point of view and construct a bounding set $\mathscr{E}_N$ for $\boldsymbol{e}^m, \boldsymbol{e}^o$ by using the fact that $\boldsymbol{e}^m, \boldsymbol{e}^o$ are linear functions of $\boldsymbol{a}_N^{true}$ and applying the energy method, then $\boldsymbol{e}^m, \boldsymbol{e}^o$ may be considered as elements of $\mathscr{E}_N$, which are independent of $\boldsymbol{a}_N^{true}$, and represent the unknown projection error. Specifically,[3] $\boldsymbol{e}^m$ represents the error of projecting the differential operator $\mathcal{A}$ and the second equation in (3.1) is necessary to filter out inadmissible $\boldsymbol{e}^m$, $\boldsymbol{e}^o$. From this standpoint, the DAE (3.1) serves as grounds for deriving a robust estimate for $\boldsymbol{a}_N^{true}$ and so every term in
(3.1) provides information which is then used in the actual numerical computations.

The proof of Proposition 3.1 relies on three lemmas, the proofs of which are technical and not needed in further calculations. They are provided in the appendix. We make use of the following definitions: define $\rho_1(\boldsymbol{x}, t) := \|\boldsymbol{v}(\boldsymbol{x}, t)\|^2_{\mathbb{R}^n}$, $\rho_2(\boldsymbol{x}, t) := \|J_{\boldsymbol{v}}(\boldsymbol{x}, t)\|_2^2$ and set

$$\begin{aligned} \mu_1(t) &:= \|\rho_1(\cdot, t)\|_{L^\infty(\Omega)} + 2\|\lambda_1^{-1}\rho_2(\cdot, t) + \rho_1(\cdot, t)\|_{L^\infty(\Omega)}\,, \\ C(\varepsilon, \boldsymbol{v}) &= \frac{9}{\varepsilon^2}\left(1 + \frac{2}{\varepsilon}\int_0^T \|\rho_1(\cdot, t)\|_{L^\infty(\Omega)} \exp\left\{\frac{2}{\varepsilon}\int_0^t \|\rho_1(\cdot, s)\|_{L^\infty(\Omega)} ds\right\} dt\right)\,, \\ S^{-1} &:= \|\mu_1\|_{L^\infty(0,T)} C(\varepsilon, \boldsymbol{v}) \max\{\underline{q}_0^{-1}, \max_t \underline{q}^{-1}\}\,, \\ V^{-1} &:= \sum_{k=1}^M \|(\mathcal{I} - \mathcal{P}_N^\dagger \mathcal{P}_N) g_k\|^2_{L^2(0,T,L^2(\Omega))} C(\varepsilon, \boldsymbol{v}) \max\{\underline{q}_0^{-1}, \max_t \underline{q}^{-1}\}\,. \end{aligned}$$

Since Ω may have a nonsmooth boundary (for instance, a rectangular domain used in the case study), we need to assure that (2.1) has a unique solution I such that $\Delta I(\cdot, t)$ is well defined for almost all $t \in (0, T)$. This is demonstrated by the following lemma.

[3] We refer the reader to subsection 3.2, where the detailed interpretation of $\boldsymbol{e}^m$, $\boldsymbol{e}^o$ is provided.

LEMMA 3.2. *Equation* (2.1) *has a unique solution* $I(\cdot,t) \in H_0^1(\Omega)$ *such that* $\Delta I(\cdot,t) \in L^2(\Omega)$ *for almost all* $t \in (0,T)$*, provided* $f \in L^2(0,T,L^2(\Omega))$*,* $I_0 \in H^2(\Omega) \cap H_0^1(\Omega)$*, and* Ω *is a convex bounded open domain.*

We also require the following estimate for $\boldsymbol{e}^o \cdot \boldsymbol{e}^o$.

LEMMA 3.3.

$$\boldsymbol{e}^o \cdot \boldsymbol{e}^o \le 2\lambda_{N+1}^{-1} \|\lambda_1^{-1}\rho_2(\cdot,t) + \rho_1(\cdot,t)\|_{L^\infty(\Omega)} \|\Delta I(\cdot,t)\|_{L^2(\Omega)}^2 . \tag{3.3}$$

Finally, we require a bound on $\|\Delta I(\cdot,t)\|_{L^2(\Omega)}^2$.

LEMMA 3.4.

$$\|\Delta I\|_{L^2(0,T,L^2(\Omega))}^2 \le C(\varepsilon,\boldsymbol{v})(\|\nabla I_0\|_{L^2(\Omega)}^2 + \|f\|_{L^2(0,T,L^2(\Omega))}^2) . \tag{3.4}$$

Now we turn to the proof of Proposition 3.1.

Proof. Formally, the first claim is almost obvious. Indeed, let us define

$$e(\boldsymbol{x},t) := \mathcal{A}\mathcal{P}_N^\dagger \mathcal{P}_N I(\boldsymbol{x},t) - \mathcal{P}_N^\dagger \mathcal{P}_N \mathcal{A} I(\boldsymbol{x},t) . \tag{3.5}$$

Since $\boldsymbol{a}_N^{true}(t) = \mathcal{P}_N I(\cdot,t)$, we show that $\boldsymbol{a}_N^{true}$ solves

$$\partial_t \mathcal{P}_N^\dagger \boldsymbol{a} = \mathcal{P}_N^\dagger \mathcal{P}_N \partial_t I = -\mathcal{A}\mathcal{P}_N^\dagger \boldsymbol{a} + e + \mathcal{P}_N^\dagger \mathcal{P}_N f . \tag{3.6}$$

Multiplying (3.6) by $\mathcal{P}_N$ and noting $\mathcal{P}_N\mathcal{P}_N^\dagger = \mathcal{I}$, we find that $\boldsymbol{a}_N^{true}$ solves the first equation in (3.1) for $\boldsymbol{e}^m = \mathcal{P}_N e$. On the other hand, (3.6) has a solution if and only if $-\mathcal{A}\mathcal{P}_N^\dagger \boldsymbol{a} + e$ is in the range of $\mathcal{P}_N^\dagger$. This holds true, in turn, if $(\mathcal{I} - \mathcal{P}_N^\dagger\mathcal{P}_N)\mathcal{A}\mathcal{P}_N^\dagger \boldsymbol{a} = (\mathcal{I} - \mathcal{P}_N^\dagger\mathcal{P}_N)e$. By (3.5), $(\mathcal{I} - \mathcal{P}_N^\dagger\mathcal{P}_N)e(t) = (\mathcal{I} - \mathcal{P}_N^\dagger\mathcal{P}_N)\mathcal{A}\mathcal{P}_N^\dagger \boldsymbol{a}_N^{true}$, and, recalling that $(\mathcal{P}_N^\dagger)' = \mathcal{P}_N$, we compute

$$\|(\mathcal{I} - \mathcal{P}_N^\dagger\mathcal{P}_N)\mathcal{A}\mathcal{P}_N^\dagger \boldsymbol{a}_N^{true}\|_{L^2(\Omega)}^2 = (S_N - A_N' A_N)\boldsymbol{a}_N^{true} \cdot \boldsymbol{a}_N^{true} = \|H_N \boldsymbol{a}_N^{true}\|_{\mathbb{R}^N}^2 . \tag{3.7}$$

Thus, $\boldsymbol{a}_N^{true}$ solves the second equation in (3.1) for $\boldsymbol{e}^o = -H_N \boldsymbol{a}_N^{true}$. To see that the third equation in (3.1) holds for $\boldsymbol{a}_N^{true}$, it is sufficient to set $\boldsymbol{w} = (v_1 \dots v_M)'$, where $v_k(t) := \langle g_k(\cdot,t), (\mathcal{I} - \mathcal{P}_N^\dagger\mathcal{P}_N)I(\cdot,t)\rangle_{L^2(\Omega)}$.

Let us prove that $\boldsymbol{e}^m$, $\boldsymbol{e}^o$ and $\boldsymbol{w}$ satisfy (3.2). In order to estimate $\boldsymbol{e}^m \cdot \boldsymbol{e}^m$, we recall that $I^N := \mathcal{P}_N^\dagger \mathcal{P}_N I$ and so $\boldsymbol{e}^m \cdot \boldsymbol{e}^m = \|\mathcal{P}_N \mathcal{A}(I^N - I)\|_{\mathbb{R}^N}^2$. Let us compute $\mathcal{A}(I^N - I)$. Noting that, by Lemma 3.2, $\Delta I(\cdot,t) \in L^2(\Omega)$, we write

$$-\Delta I(\boldsymbol{x},t) = \sum_{i\in\mathbb{N}} \langle \varphi_i, -\Delta I(\cdot,t)\rangle_{L^2(\Omega)} \varphi_i(\boldsymbol{x}) = \sum_{i\in\mathbb{N}} \lambda_i a_i(t)\varphi_i(\boldsymbol{x}) . \tag{3.8}$$

To prove this, we apply an integration by parts formula [9, p. 52] (we omit the argument $(\boldsymbol{x},t)$ below to make the notation more convenient):

$$\int_\Omega (\partial_{x_i} u) v d\boldsymbol{x} + \int_\Omega u(\partial_{x_i} v) d\boldsymbol{x} = \int_{\partial\Omega} \mathrm{tr}(u)\,\mathrm{tr}(v)\nu_i d\sigma \,, \forall\, u, v \in H^1(\Omega) , \tag{3.9}$$

where $\mathrm{tr}(u)$ denotes the trace of u on $\partial\Omega$ and $\boldsymbol{\nu} = (\nu_1 \dots \nu_n)'$ denotes the outward pointing normal vector for $\partial\Omega$. Namely, recalling that $I(\cdot,t) \in H_0^1(\Omega)$ by Lemma 3.2, $\varphi \in H_0^1(\Omega)$ by definition, and $u \in H_0^1(\Omega) \Leftrightarrow \mathrm{tr}(u) = 0$ (see, for instance, [9, p. 39]), we integrate $\langle \varphi_i, \Delta I(\cdot,t)\rangle_{L^2(\Omega)}$ by parts twice to get $\langle \varphi_i, -\Delta I(\cdot,t)\rangle_{L^2(\Omega)} =$

$\langle -\Delta\varphi_i, I(\cdot,t)\rangle_{L^2(\Omega)} = \lambda_i a_i(t)$. Now, the orthogonality condition $\langle\varphi_k,\varphi_s\rangle_{L^2(\Omega)} = \delta_{ks}$ and (3.8) imply that for almost all $t \in (0,T)$,

$$\langle \Delta I(\cdot,t), \Delta I(\cdot,t)\rangle_{L^2(\Omega)} = \sum_{i\in\mathbb{N}} \lambda_i^2 a_i^2(t) < +\infty\,, \tag{3.10}$$

and so $-\Delta(I - I^N) = \sum_{i>N} \lambda_i a_i \varphi_i$ for $I^N = \sum_{i\le N} a_i\varphi_i$. Combining the obtained representation with the orthogonality condition $\langle\varphi_k,\varphi_s\rangle_{L^2(\Omega)} = 0$ for $k \le N < s$, we get that $\langle\varphi_k, -\Delta(I - I^N)\rangle^2_{L^2(\Omega)} = 0$ and so

$$\langle\varphi_k, \mathcal{A}(I - I^N)\rangle_{L^2(\Omega)} = \langle\varphi_k, \boldsymbol{v}\cdot\nabla(I - I^N)\rangle_{L^2(\Omega)}\ \forall k \le N\,. \tag{3.11}$$

Now, recalling that $\|\varphi_k\|^2_{L^2(\Omega)} = 1$ and $\|\boldsymbol{v}\cdot\nabla(I - I^N)\|^2_{L^2(\Omega)} = \sum_{k=1}^{\infty}\langle\varphi_k, \boldsymbol{v}\cdot\nabla(I - I^N)\rangle^2_{L^2(\Omega)}$, we estimate $\boldsymbol{e}^m\cdot\boldsymbol{e}^m$:

$$\begin{aligned}
\boldsymbol{e}^m\cdot\boldsymbol{e}^m &= \|\mathcal{P}_N\mathcal{A}(I^N - I)\|^2_{\mathbb{R}^N} = \sum_{k=1}^{N}\langle\varphi_k, \boldsymbol{v}\cdot\nabla(I - I^N)\rangle^2_{L^2(\Omega)}\\
&\le \|\rho_1(\cdot,t)\|_{L^\infty(\Omega)}\|\nabla(I - I^N)\|^2_{L^2(\Omega)}\\
&= \|\rho_1(\cdot,t)\|_{L^\infty(\Omega)}\langle -\Delta(I - I^N), I - I^N\rangle_{L^2(\Omega)}\\
&= \|\rho_1(\cdot,t)\|_{L^\infty(\Omega)}\sum_{i>N}\lambda_i a_i^2(t) \le \|\rho_1(\cdot,t)\|_{L^\infty(\Omega)}\lambda_{N+1}^{-1}\sum_{i>N}\lambda_i^2 a_i^2(t)\\
&\le \|\rho_1(\cdot,t)\|_{L^\infty(\Omega)}\lambda_{N+1}^{-1}\|\Delta I(\cdot,t)\|^2_{L^2(\Omega)}\,.
\end{aligned} \tag{3.12}$$

An estimate for $\boldsymbol{e}^o\cdot\boldsymbol{e}^o$ is provided by Lemma 3.3.

Let us estimate $\boldsymbol{w}$. Define $g_k^\perp := (\mathcal{I} - \mathcal{P}_N^\dagger\mathcal{P}_N)g_k$. Recalling the definition of v_k given above, we compute $v_k(t) = \langle g_k, (\mathcal{I} - \mathcal{P}_N^\dagger\mathcal{P}_N)I(\cdot,t)\rangle_{L^2(\Omega)} = \sum_{s>N}\langle g_k,\varphi_s\rangle_{L^2(\Omega)}a_s(t)$ and so, by applying the Cauchy–Schwarz–Bunyakovsky inequality, we obtain

$$\begin{aligned}
v_k(t) &= \sum_{s>N}\langle g_k,\varphi_s\rangle_{L^2(\Omega)}a_s(t) \le \left(\sum_{s>N}\lambda_s^{-2}\langle g_k,\varphi_s\rangle^2_{L^2(\Omega)}\right)^{\frac12}\left(\sum_{s>N}\lambda_s^2 a_s^2(t)\right)^{\frac12}\\
&\le \lambda_{N+1}^{-1}\|g_k^\perp\|_{L^2(\Omega)}\|\Delta I(\cdot,t)\|_{L^2(\Omega)}\,.
\end{aligned} \tag{3.13}$$

Let us note that, by the definition of Q_0, Q, we have

$$\|\nabla I_0\|^2_{L^2(\Omega)} + \|f\|^2_{L^2(0,T,L^2(\Omega))} \le \max\{\underline{q}_0^{-1}, \max_t \underline{q}^{-1}\}\,. \tag{3.14}$$

Now, by integrating (3.13) and using (3.4) from Lemma 3.4 followed by (3.14) and the definition of V, we get $\int_0^T v_k^2(t)\,dt \le \lambda_{N+1}^{-2}\|g_k^\perp\|^2_{L^2(0,T,L^2(\Omega))}C(\varepsilon,\boldsymbol{v})\max\{\underline{q}_0^{-1}, \max_t \underline{q}^{-1}\}$ and so

$$\int_0^T \lambda_{N+1}\|V^{\frac12}\boldsymbol{w}\|^2_{\mathbb{R}^M}\,dt \le \lambda_{N+1}^{-1}\,. \tag{3.15}$$

To conclude the proof, we note that

$$\|\nabla I_0\|^2_{L^2(\Omega)} \le \underline{q}_0^{-1}\|Q_0^{\frac12}\nabla I_0\|^2_{L^2(\Omega)} \text{ and } \|f\|^2_{L^2(0,T,L^2(\Omega))} \le \max_t \underline{q}^{-1}\|Q^{\frac12}f\|^2_{L^2(0,T,L^2(\Omega))}\,.$$

Finally, to get (3.2) we add (3.12) to (3.3) from Lemma 3.3, integrate from 0 to T, and bound the right-hand side of the resulting inequality by using the definition of μ_1, (3.4), and (3.14). Then we multiply the resulting inequality by $S\lambda_{N+1}^{-\frac12}$ and add the result to the sum of (2.2) and (3.15). This completes the proof. ∎

3.2. Comparison to the classical Galerkin approach. In this section we briefly contrast the extended Galerkin formulation of the previous section with the classical Galerkin method. To simplify the presentation, we assume for a moment that $f = 0$ in (2.1). The classical Galerkin projection approach is built upon the following requirement [12, p. 43]:

$$\frac{dI^N}{dt} + \mathcal{A}I^N \perp \operatorname{span}\{\varphi_1 \dots \varphi_N\}\,, \tag{3.16}$$

where $I^N = \mathcal{P}_N^\dagger \mathcal{P}_N I = \sum_{i=1}^N a_i \varphi_i$ approximates I solving (2.1). This condition yields the following ODE for determining $\boldsymbol{a} = (a_1 \dots a_N)'$:

$$\frac{d\boldsymbol{a}}{dt} = -A_N \boldsymbol{a}, \quad \boldsymbol{a}(0) = \mathcal{P}_N I_0\,. \tag{3.17}$$

Let us investigate the connection between (3.1) and (3.17). We note that the basic assumption (3.16) of the Galerkin method holds true for $\boldsymbol{a}_N^{true}$ if and only if

$$\frac{d\mathcal{P}_N^\dagger \boldsymbol{a}_N^{true}}{dt} + A\mathcal{P}_N^\dagger \boldsymbol{a}_N^{true} \perp \operatorname{span}\{\varphi_1 \dots \varphi_N\}\,.$$

Now, by (3.6), the latter is true if and only if $\mathcal{P}_N e(t) = 0$. Recalling (3.5), we rewrite e as follows:

$$e = (\mathcal{I} - \mathcal{P}_N^\dagger \mathcal{P}_N)\mathcal{A}\mathcal{P}_N^\dagger \mathcal{P}_N I + \mathcal{P}_N^\dagger \mathcal{P}_N \mathcal{A}(\mathcal{P}_N^\dagger \mathcal{P}_N - \mathcal{I})I\,. \tag{3.18}$$

Now, we compute $\mathcal{P}_N e(t) = \mathcal{P}_N \mathcal{A}(\mathcal{P}_N^\dagger \mathcal{P}_N - \mathcal{I})I$ and so $\boldsymbol{a}(t) = \boldsymbol{a}_N^{true}$ if and only if $\mathcal{P}_N \mathcal{A}(\mathcal{P}_N^\dagger \mathcal{P}_N - \mathcal{I})I = 0$. In other words, $\boldsymbol{a}(t) = \boldsymbol{a}_N^{true}(t)$ if the $\mathcal{A}$-image of the projection error $(\mathcal{P}_N^\dagger \mathcal{P}_N - \mathcal{I})I$ is orthogonal to the span of $\{\varphi_k\}_{k=1}^N$ and, therefore, has no impact on the dynamics of $\boldsymbol{a}_N^{true}$. We stress that $\mathcal{P}_N e(t) \neq 0$ in the general case but there are important special cases when this holds true, namely, $e = 0$ provided that $\mathcal{A}\varphi_k = \alpha_k \varphi_k$. This suggests the following interpretation for (3.18): the norm of e quantifies the degree to which the subspace generated by $\{\varphi_k\}_{k=1}^N$ differs from an eigenspace of $\mathcal{A}$. More generally, $\mathcal{P}_N e = 0$ if $\mathcal{P}_N^\dagger \mathcal{P}_N$ commutes with $\mathcal{A}$. In this case, (3.17) gives a closed system for $\boldsymbol{a}_N^{true}$: it contains all the required information to describe how the exact projection coefficients evolve over time. We emphasize that, in practice, the assumption $\mathcal{P}_N e = 0$ is not easy to check (for a given set of basis functions), as e depends on the solution I which is unknown. Therefore, in practice, the Galerkin system (3.17) is usually nonclosed.

In contrast to the classical Galerkin method, the solution proposed by Proposition 3.1 is to consider $\boldsymbol{e}^m = \mathcal{P}_N e$ as an unknown deterministic input for (3.17) and construct an a priori estimate for $\boldsymbol{e}^m = \mathcal{P}_N e$ in the form (3.2) using information about the coefficients of $\mathcal{A}$, domain Ω, and data I_0, f. As a result, the true coefficients $\boldsymbol{a}_N^{true}$ belong to the set of solutions of (3.1). The information provided by the second equation in (3.1) allows filtering out inadmissible $\boldsymbol{e}^m$. In fact, it bounds the norm of $(\mathcal{I} - \mathcal{P}_N^\dagger \mathcal{P}_N)\mathcal{A}\mathcal{P}_N^\dagger \hat{\boldsymbol{a}}_N$, representing the energy of the $\mathcal{A}$-image of the projected solution $\mathcal{P}_N^\dagger \hat{\boldsymbol{a}}_N$ in the orthogonal complement of $\operatorname{span}\{\varphi_1 \dots \varphi_N\}$. This allows one, in turn, to narrow down the set of all admissible $\boldsymbol{a}$ solving (3.1). Finally, the resulting DAE (3.1) together with ellipsoid (3.2) represents a closed system for $\boldsymbol{a}_N^{true}$.

3.3. Minimax projection coefficients. In this section we construct the minimax approximation of the solution of the DAE (3.1). This solution minimizes the maximum error over the parameter set $\mathscr{E}_N$ containing the true projection coefficients. Subsequently, we show that the minimax solution can be obtained as the solution of an equivalent optimal control problem. Finally, we cast the optimal control problem in a form that facilitates its numerical solution.

Following the definition of the minimax estimate given in section 2, we will be looking for an estimate of a linear function $\boldsymbol{\ell} \cdot \boldsymbol{a}(T)$ of the state of (2.7) within the class of linear functionals:

$$\boldsymbol{u}(\boldsymbol{y}) = \int_0^T \boldsymbol{u}(t) \cdot \boldsymbol{y}(t)\, dt\,, \boldsymbol{u} \in L^2(0,T,\mathbb{R}^M)\,.$$

DEFINITION 3.5. *A linear estimate* $\hat{\boldsymbol{u}}(\boldsymbol{y}) = \int_0^T \hat{\boldsymbol{u}} \cdot \boldsymbol{y}\, dt$ *is called a* minimax estimate *if* $\inf_{\boldsymbol{u}} \sigma(\boldsymbol{u}, T, \boldsymbol{\ell}) = \sigma(\hat{\boldsymbol{u}}, T, \boldsymbol{\ell})$*, where*

$$\sigma(\boldsymbol{u}, T, \boldsymbol{\ell}) := \sup_{(I_0, f, \boldsymbol{e}^m, \boldsymbol{e}^o, \boldsymbol{w}) \in \mathscr{E}_N, \boldsymbol{e}} \mathbb{E}(\boldsymbol{\ell} \cdot \boldsymbol{a}(T) - \boldsymbol{u}(\boldsymbol{y}))^2\,. \tag{3.19}$$

The number $\hat{\sigma}(T, \boldsymbol{\ell}) = \sigma(\hat{\boldsymbol{u}}, T, \boldsymbol{\ell})$ *is called a* minimax error*.*

In fact, $\sigma(\boldsymbol{u}, T, \boldsymbol{\ell})$ can be interpreted as yielding the "worst" realization of the unknown deterministic parameters satisfying (3.2) and covariance operator of $\boldsymbol{e}$ satisfying (2.4). Since $\hat{\boldsymbol{u}}$ possesses minimal worst-case error $\hat{\sigma}$, it follows that $\hat{\boldsymbol{u}}$ is robust with respect to any realization of unknown parameters.

Following [26], we apply the generalized Kalman duality principle to construct the minimax estimate $\hat{\boldsymbol{u}}$ for the DAE (3.1). Define

$$Q_{0N} := \{\lambda_k^{-1}\lambda_s^{-1}\langle Q_0 \nabla\varphi_k, \nabla\varphi_s\rangle_{L^2(\Omega)}\}_{k,s=1}^N$$

and set $Q_N := (\mathcal{P}_N Q \mathcal{P}_N^\dagger)^{-1} + \lambda_{N+1}^{-\frac{1}{2}} S^{-1}\mathcal{I}$, $R_N := \frac{1}{\mu_N} R^{-1} + \lambda_{N+1}^{-1} V^{-1}\mathcal{I}$.

PROPOSITION 3.6. *The minimax estimate* $\hat{\boldsymbol{u}}$ *is the unique minimum point of* $\sigma(\boldsymbol{u}, T, \boldsymbol{\ell})$*, where*

$$\begin{aligned} \frac{1}{\mu_N}\sigma(\boldsymbol{u}, T, \boldsymbol{\ell}) &= \min_{\boldsymbol{g}} Q_{0,N}^{-1}\boldsymbol{z}(t_0)\cdot\boldsymbol{z}(t_0) \\ &\quad + \int_0^T (Q_N \boldsymbol{z}\cdot\boldsymbol{z} + R_N\boldsymbol{u}\cdot\boldsymbol{u} + \lambda_{N+1}^{-\frac{1}{2}} S^{-1}\boldsymbol{g}\cdot\boldsymbol{g})\, dt, \end{aligned} \tag{3.20}$$

$$\frac{d\boldsymbol{z}}{dt} = A_N'\boldsymbol{z} - H_N'\boldsymbol{g} + C_N'\boldsymbol{u}\,, \boldsymbol{z}(T) = \boldsymbol{\ell}\,. \tag{3.21}$$

Proof. Let us compute $\sigma(\boldsymbol{u}, T, \boldsymbol{\ell})$. By recalling the third equation in (3.1), we compute $\boldsymbol{u}(\boldsymbol{y}) = \langle \boldsymbol{u}, C_N\boldsymbol{a} + \boldsymbol{w}\rangle_{L^2(0,T)} + \langle \boldsymbol{u}, \boldsymbol{e}\rangle_{L^2(0,T)}$. Combining this with the assumption $\mathbb{E}\boldsymbol{e} = 0$, we get

$$\mathbb{E}(\boldsymbol{\ell}\cdot\boldsymbol{a}(T) - \boldsymbol{u}(\boldsymbol{y}))^2 = \mathbb{E}\langle \boldsymbol{u}, \boldsymbol{e}\rangle^2_{L^2(0,T)} + \left(\boldsymbol{\ell}\cdot\boldsymbol{a}(T) - \langle \boldsymbol{u}, C_N\boldsymbol{a} + \boldsymbol{w}\rangle_{L^2(0,T)}\right)^2\,.$$

Clearly, for any $\boldsymbol{u} \in L^2(0,T)$, we can find at least one $\boldsymbol{z}$, $\boldsymbol{g}$ such that $\boldsymbol{z}$, $\boldsymbol{g}$, and $\boldsymbol{u}$ satisfy the adjoint equation (3.21). Using this observation and integrating by parts the term $\langle \boldsymbol{u}, C_N\boldsymbol{a} + \boldsymbol{w}\rangle_{L^2(0,T)}$, we find

$$\begin{aligned} \sigma(\boldsymbol{u}, T, \boldsymbol{\ell}) &= \sup_{(I_0, f, \boldsymbol{e}^m, \boldsymbol{e}^o, \boldsymbol{w}) \in \mathscr{E}_N} \alpha^2 + \sup_{\boldsymbol{e}} \mathbb{E}\langle \boldsymbol{u}, \boldsymbol{e}\rangle^2_{L^2(0,T)}\,, \\ \alpha &:= \langle I_0, \mathcal{P}_N^\dagger \boldsymbol{z}(t_0)\rangle_{L^2(\Omega)} + \int_0^T (\langle f, \mathcal{P}_N^\dagger \boldsymbol{z}\rangle_{L^2(\Omega)} + \boldsymbol{e}^m\cdot\boldsymbol{z} + \boldsymbol{e}^o\cdot\boldsymbol{g} - \boldsymbol{u}\cdot\boldsymbol{w})\, dt\,. \end{aligned} \tag{3.22}$$

By the Cauchy–Schwarz–Bunyakovsky inequality, we get

$$\mathbb{E}\langle \boldsymbol{u}, \boldsymbol{e}\rangle^2_{L^2(0,T)} \le \langle R^{-1}\boldsymbol{u}, \boldsymbol{u}\rangle^2_{L^2(0,T)} \mathbb{E}\langle R\boldsymbol{e}, \boldsymbol{e}\rangle^2_{L^2(0,T)}\,.$$

Noting that $\mathbb{E}\langle R\boldsymbol{e}, \boldsymbol{e}\rangle_{L^2(0,T)} = \int_0^T \operatorname{trace}(R\,\mathbb{E}\,\boldsymbol{e}'(t)\boldsymbol{e}(t))\,dt$ and recalling (2.4), we obtain

$$\sup_{\boldsymbol{e}} \mathbb{E}\langle \boldsymbol{u}, \boldsymbol{e}\rangle^2_{L^2(0,T)} = \langle R^{-1}\boldsymbol{u}, \boldsymbol{u}\rangle^2_{L^2(0,T)}\,. \tag{3.23}$$

Let us estimate $\sup_{\mathscr{E}_N} \alpha$. We first note that

$$\begin{aligned}\langle I_0, \mathcal{P}_N^\dagger \boldsymbol{z}(t_0)\rangle_{L^2(\Omega)} &= \sum_{k=1}^N \langle I_0, \varphi_k\rangle_{L^2(\Omega)} z_k(t_0) = -\sum_{k=1}^N \lambda_k^{-1}\langle I_0, \Delta\varphi_k\rangle_{L^2(\Omega)} z_k(t_0)\\ &= \sum_{k=1}^N \lambda_k^{-1}\langle \nabla I_0, \nabla\varphi_k\rangle_{L^2(\Omega)} z_k(t_0)\,.\end{aligned}$$

Now, by using the latter representation and (3.2), we compute $\sup_{\mathscr{E}_N} \alpha$ by applying the generalized Cauchy–Schwarz–Bunyakovsky inequality:

$$\frac{1}{\mu_N}\sup_{I_0, f, \boldsymbol{e}^m, \boldsymbol{e}^o, \boldsymbol{w}} \alpha^2 = Q_{0N}^{-1}\boldsymbol{z}(t_0)\cdot\boldsymbol{z}(t_0) + \int_0^T Q_N\boldsymbol{z}\cdot\boldsymbol{z} + \lambda_{N+1}^{-\frac{1}{2}} S^{-1}\boldsymbol{g}\cdot\boldsymbol{g} + V^{-1}\lambda_{N+1}^{-1}\boldsymbol{u}\cdot\boldsymbol{u}\,dt\,. \tag{3.24}$$

Combining this with (3.23) and recalling (3.22), we find that $\frac{1}{\mu_N}\sigma(\boldsymbol{u}, T, \boldsymbol{\ell})$ is represented by (3.24). Now, we note that $\boldsymbol{z}$ is uniquely defined by $\boldsymbol{g}$ and $\boldsymbol{u}$ through (3.21) and $\boldsymbol{g}$ may be considered as a "free parameter" which belongs to the "null-space" of the linear operator associated to (3.21). (See [22] for further details.) In other words, the adjoint DAE (3.21) is overdetermined as the original DAE (3.1) is underdetermined. Since the minimax estimate $\hat{\boldsymbol{u}}$ should have the minimal worst-case estimation error, the latter is represented by σ and σ depends on the "free parameter" $\boldsymbol{g}$; it follows that we can determine the minimax estimate $\hat{\boldsymbol{u}}$ by minimizing σ with respect to $\boldsymbol{u}$, $\boldsymbol{g}$, provided $\boldsymbol{z}$ solves the adjoint equation (3.21). This completes the proof. ∎

COROLLARY 3.7. *The unique solution of* (3.20) *is given by* $\hat{\boldsymbol{u}} = R_N^{-1}C_N V U^{-1}(T)\boldsymbol{\ell}$ *and* $\hat{\boldsymbol{g}} = -\lambda_{N+1}^{\frac{1}{2}} S H_N V U^{-1}(T)\boldsymbol{\ell}$. *The optimal value of the cost is* $\sigma(\boldsymbol{u}, T, \boldsymbol{\ell}) = \mu_N K_N(T)\boldsymbol{\ell}\cdot\boldsymbol{\ell}$, *provided* $K_N = VU^{-1}$ *and the matrix-valued functions* V, U *solve the following linear Hamiltonian ODE:*

$$\begin{aligned}&\dot U = A_N' U + (\lambda_{N+1}^{\frac{1}{2}} S H_N' H_N + C_N' R_N^{-1} C_N) V\,,\\ &\dot V = -A_N V + Q_N U\,, V(t_0) = Q_{0N}^{-1}\,, U(0) = \mathcal{I}.\end{aligned} \tag{3.25}$$

The minimax estimate $\hat{\boldsymbol{u}}(\boldsymbol{y})$ *may be represented as an output of the linear system, that is,* $\hat{\boldsymbol{u}}(\boldsymbol{y}) = \boldsymbol{\ell}\cdot\hat{\boldsymbol{a}}_N(T)$, *where* $\hat{\boldsymbol{a}}_N$ *solves the following ODE:*

$$\begin{aligned}\frac{d\hat{\boldsymbol{a}}_N}{dt} =& - A_N\hat{\boldsymbol{a}}_N - K_N\left(\lambda_{N+1}^{\frac{1}{2}} S H_N' H_N + C_N' R_N^{-1} C_N\right)\hat{\boldsymbol{a}}_N\\ &+ K_N C_N' R_N^{-1}\boldsymbol{y}\,, \quad \hat{\boldsymbol{a}}_N(0) = 0\,.\end{aligned} \tag{3.26}$$

In particular, we have that for all $\boldsymbol{\ell}$, $\mathbb{E}(\boldsymbol{\ell}\cdot\boldsymbol{a}_N^{true}(t) - \boldsymbol{\ell}\cdot\hat{\boldsymbol{a}}_N(t)))^2 \le \mu_N K_N(t)\boldsymbol{\ell}\cdot\boldsymbol{\ell}$.

Proof. The proof of the first part of the Corollary follows from the well-known results of linear control theory [19]. The existence of U^{-1} follows from [19, p. 121,

L.4.1]. Now, to prove that $\hat{\boldsymbol{u}}(\boldsymbol{y}) = \boldsymbol{\ell} \cdot \hat{\boldsymbol{a}}_N(T)$, one needs to use the representation $\hat{\boldsymbol{u}}(t) = R_N^{-1} C_N(t) V(t) U^{-1}(T) \boldsymbol{\ell}$, (3.25)–(3.26), and integration by parts. Detailed derivation for DAEs may be found in [26]. ∎

Remark 2. Let us note that $\lim_{k\to\infty} \frac{\lambda_k^{\frac{n}{2}}}{k} = \frac{(2\pi)^n}{\alpha(n)|\Omega|}$ by Weyl's Law [7, p. 356], where $|\Omega|$ denotes the volume of Ω and $\alpha(n)$ is the volume of the unit ball in $\mathbb{R}^n$. This observation makes it clear that $Q_N = (\mathcal{P}_N Q \mathcal{P}_N^\dagger)^{-1} + \lambda_{N+1}^{-\frac{1}{2}} S^{-1} \mathcal{I} \approx (\mathcal{P}_N Q \mathcal{P}_N^\dagger)^{-1}$ and $R_N^{-1} = (\frac{1}{\mu_N} R^{-1} + \lambda_{N+1}^{-1} V^{-1} \mathcal{I})^{-1} \approx R$ for large enough N. On the other hand, $\lambda_{N+1}^{\frac{1}{2}} S H_N' H_N \to 0$ by (A.3), (3.3), and so, for large enough N, we have that U, V can be made arbitrarily close to the solutions of the system

$$
\begin{aligned}
&\dot{U} = A_N' U + C_N' R C_N V \,, U(0) = \mathcal{I} \,, \\
&\dot{V} = -A_N V + (\mathcal{P}_N Q \mathcal{P}_N^\dagger)^{-1} U \,, V(t_0) = Q_{0N}^{-1}
\end{aligned} \tag{3.27}
$$

and the minimax filter $\hat{\boldsymbol{a}}_N$ becomes arbitrarily close to the solution of $\frac{d\hat{\boldsymbol{a}}_N}{dt} = -A_N \hat{\boldsymbol{a}}_N - K_N C_N' R C_N \hat{\boldsymbol{a}}_N + K_N C_N' R \boldsymbol{y}$, $\hat{\boldsymbol{a}}_N(0) = 0$. In other words, the minimax projection method guarantees that (3.26) converges to the minimax estimate of the infinite dimensional system (2.1). On the other hand, the constant $C(\varepsilon, \boldsymbol{v})$ in the estimate (3.4) for the Laplacian $\Delta I(\boldsymbol{x}, t)$ is very conservative: the errors $\boldsymbol{e}^m, \boldsymbol{e}^o$ decay faster than $\lambda_{N+1}^{-1} \|\Delta I(\cdot, t)\|^2_{L^2(\Omega)}$, as can be seen from (3.12) and (A.3). Therefore, in practice, it is not necessary (but, of course, is sufficient) to choose N so that $\lambda_{N+1}^{-\frac{1}{2}} S^{-1}$, $\lambda_{N+1}^{-1} V^{-1}$ become negligible when compared with $(\mathcal{P}_N Q \mathcal{P}_N^\dagger)^{-1}$ and R. A practical way to choose N would be to make sure that $\mu_N \approx 1$, $\|H_N' H_N\| \approx 0$, $\|(\mathcal{I} - \mathcal{P}_N^\dagger \mathcal{P}_N) g_k\|_{L^2(\Omega)} \approx 0$ and N is large enough to numerically resolve the system (2.1) for the anticipated model error f. Then the estimate may be obtained from (3.27). We will apply this method in section 5.

4. Structure preserving discretization. In this section we discretize the cost function σ using a quadrature rule and compute the discrete-time minimax estimate $n \mapsto \hat{\boldsymbol{u}}_n$, which is the unique minimum point of the discretized cost. As a result, the discrete minimax estimate $n \mapsto \hat{\boldsymbol{u}}_n$ inherits the key geometric property of the continuous one: it can be represented in terms of the solution of a discrete Hamiltonian system which is, in turn, a discrete version of the continuous Hamiltonian system (3.25). This allows us, in particular, to represent $\hat{\boldsymbol{u}}_n$ in the form of discrete minimax filter $n \mapsto \hat{\boldsymbol{a}}_N^n$ and derive a representation for discrete σ which is similar to the continuous one given in Corollary 3.7. Also, we prove that the nonstationary Lyapunov function is preserved along the trajectories of $n \mapsto \hat{\boldsymbol{a}}_N^n$.

We introduce a uniform grid $t_n := nh$, $n = 1, \dots, L$, $h := \frac{T}{L}$ on $(0, T)$ and let $\{a_{ij}\}_{i,j=1}^s$, $\{b_i\}_{i=1}^s$ denote the coefficients of s-stage implicit RK method [10, p. 29] for $s \geq 1$. Now, we set $c_i := \sum_{j=1}^s a_{ij}$ and introduce the discrete cost:

$$
\begin{aligned}
\frac{1}{\mu_N} \sigma_L(\{\boldsymbol{u}_n\}, \boldsymbol{\ell}, T) = Q_{0N}^{-1} \boldsymbol{z}_0 \cdot \boldsymbol{z}_0 + h \sum_{n=0}^{L} \sum_{i=1}^{s} b_i Q_N(i, n) \boldsymbol{z}_{in} \cdot \boldsymbol{z}_{in} \\
+ h \sum_{n=0}^{L} \sum_{i=1}^{s} b_i R_N(i, n) \boldsymbol{u}_{in} \cdot \boldsymbol{u}_{in} + b_i \lambda_{N+1}^{-\frac{1}{2}} S^{-1} \boldsymbol{g}_{in} \cdot \boldsymbol{g}_{in} \,,
\end{aligned} \tag{4.1}
$$

where $R_N(i, n) := R_N(t_n + c_i h)$ and $A_N(i, n), H_N(i, n), C_N(i, n), Q_N(i, n), \boldsymbol{y}(i, n)$ are

defined analogously, and

$$\begin{aligned} &\boldsymbol{z}_{n+1} = \boldsymbol{z}_n + h\sum_{i=1}^{s} b_i \delta \boldsymbol{z}_{in}\,, \boldsymbol{z}_{in} = \boldsymbol{z}_n + h\sum_{i=1}^{s} a_{ij}\delta \boldsymbol{z}_{jn}\,, \boldsymbol{z}_L = \boldsymbol{\ell}\,,\\ &\delta \boldsymbol{z}_{in} = A'_N(i,n)\boldsymbol{z}_{in} - H'_N(i,n)\boldsymbol{g}_{in} + C'_N(i,n)\boldsymbol{u}_{in}\,. \end{aligned} \tag{4.2}$$

We set by definition $F_N(i,n) := \lambda_{N+1}^{\frac{1}{2}} S H'_N(i,n) H_N(i,n) + C'_N(i,n) R_N^{-1}(i,n) C_N(i,n)$. In the next proposition, we construct the linear discrete minimax filter $n \mapsto \hat{\boldsymbol{a}}_N^n$.

PROPOSITION 4.1. *Assume that the coefficients a_{kj}, b_j correspond to an s-stage implicit RK method of order p and $M_{jk} := b_j b_k - b_k a_{kj} - b_j a_{jk} = 0$ for $1 \le j,k \le s$, and let $\hat{\boldsymbol{a}}_N^n$ solve the discrete system*

$$\begin{aligned} &\hat{\boldsymbol{a}}_N^{n+1} = \hat{\boldsymbol{a}}_N^n + h\sum_{i=1}^{s} b_i \delta \hat{\boldsymbol{x}}_{in}\,, \hat{\boldsymbol{x}}_{in} = \hat{\boldsymbol{a}}_N^n + h\sum_{j=1}^{s} a_{ij}\delta \hat{\boldsymbol{x}}_{jn}\,,\\ &\delta \hat{\boldsymbol{x}}_{in} = -A_N(i,n)\hat{\boldsymbol{x}}_{in} - K_N^{in} F_N(i,n)\hat{\boldsymbol{x}}_{in}\\ &\qquad\qquad + K_N^{in} C'_N(i,n) R_N^{-1}(i,n)\boldsymbol{y}(i,n)\,, \hat{\boldsymbol{a}}_N(0) = 0\,, \end{aligned} \tag{4.3}$$

where $K_N^n = V_n U_n^{-1}$ and $K_N^{in} = V_{in} U_{in}^{-1}$, and U_n, V_n and U_{in}, V_{in} are defined as solutions of the variational equations:

$$\begin{aligned} &U_{n+1} = U_n + h\sum_{i=1}^{s} b_i \delta U_{in}\,, U_{in} = U_n + h\sum_{j=1}^{s} a_{ij}\delta U_{jn}\,, U_0 = \mathcal{I}\,,\\ &V_{n+1} = V_n + h\sum_{i=1}^{s} b_i \delta V_{in}\,, V_{in} = V_n + h\sum_{j=1}^{s} a_{ij}\delta V_{jn}\,, V_0 = Q_{0N}^{-1}\,,\\ &\delta U_{in} = A'_N(i,n)U_{in} + F_N(i,n)V_{in}, \delta V_{in} = -A_N(i,n)V_{in} + Q_N(i,n)U_{in}\,. \end{aligned} \tag{4.4}$$

Then $\min \sigma_L = \mu_N K_N(L)\boldsymbol{\ell}\cdot\boldsymbol{\ell}$ *and the minimax error and estimate admit the following approximation:*

$$|\hat{\sigma}(T,\boldsymbol{\ell}) - \min \sigma_L| = O(h^p)\,, |\hat{\boldsymbol{u}}(\boldsymbol{y}) - \boldsymbol{\ell}\cdot\hat{\boldsymbol{a}}_N^L| = O(h)\,. \tag{4.5}$$

Proof. Let us prove that $\min \sigma_L = \mu_N K_N(L)\boldsymbol{\ell}\cdot\boldsymbol{\ell}$. To this end, we define $\boldsymbol{z}_n := U_n U_L^{-1}\boldsymbol{\ell}$ and $\boldsymbol{p}_n := V_n \boldsymbol{z}_0$ and set $\delta \boldsymbol{z}_{in} := \delta U_{in}\boldsymbol{z}_0$, $\delta \boldsymbol{p}_{in} := \delta V_{in}\boldsymbol{z}_0$, $\boldsymbol{z}_{in} := U_{in}\boldsymbol{z}_0$ and $\boldsymbol{p}_{in} := V_{in}\boldsymbol{z}_0$. Define $\hat{\boldsymbol{u}}_n = R_N^{-1}(n)C_N(n)\boldsymbol{p}_n$ and $\hat{\boldsymbol{g}}_n = -\lambda_{N+1}^{\frac{1}{2}} S H_N(n)\boldsymbol{p}_n$, where $C_N(n)$ stands for $C_N(t_n)$ and similarly for $R_N(n)$, $H_N(n)$, with intermediate values $\hat{\boldsymbol{u}}_{in} = R_N^{-1}(i,n)C_N(i,n)\boldsymbol{p}_{in}$ and $\hat{\boldsymbol{g}}_{in} = -\lambda_{N+1}^{\frac{1}{2}} S H_N(i,n)\boldsymbol{p}_{in}$. We claim that $\hat{\boldsymbol{u}}_n$ and $\hat{\boldsymbol{g}}_n$ minimize the discrete cost function σ_L defined by (4.1) over solutions of (4.2). To see this, one needs to check that $\sigma_L(\{\boldsymbol{u}_n\},\boldsymbol{\ell},T) - \sigma_L(\{\hat{\boldsymbol{u}}_n\},\boldsymbol{\ell},T) \ge 0$ for any $\{\boldsymbol{u}_n,\boldsymbol{g}_n\}$. The latter can be proved by plugging the expressions for $\boldsymbol{p}_n$, $\boldsymbol{z}_n$ into the right-hand side of the following subgradient inequality,

$$\frac{1}{\mu_N}\sigma_L(\{\boldsymbol{u}_n\},\boldsymbol{\ell},T) - \frac{1}{\mu_N}\sigma_L(\{\hat{\boldsymbol{u}}_n\},\boldsymbol{\ell},T) \ge \nabla_{\{\boldsymbol{u}_n,\boldsymbol{g}_n\}}\sigma_L(\{\hat{\boldsymbol{u}}_n\},\boldsymbol{\ell},T)\cdot\sum_{n=0}^{L}\begin{pmatrix}\boldsymbol{u}_n-\hat{\boldsymbol{u}}_n\\ \boldsymbol{g}_n-\hat{\boldsymbol{g}}_n\end{pmatrix}$$

and integrating the resulting expression by parts using the formula

$$\boldsymbol{z}_L\cdot\boldsymbol{p}_L - \boldsymbol{z}_0\cdot\boldsymbol{p}_0 = h\sum_{n=0}^{L}\sum_{i=1}^{s} b_i \delta\boldsymbol{z}_{in}\cdot\boldsymbol{p}_{in} + b_i\boldsymbol{z}_{in}\cdot\delta\boldsymbol{p}_{in}\,, \tag{4.6}$$

which holds for any RK method satisfying $M_{jk} \equiv 0$. Now, recalling the definitions of $\boldsymbol{p}_n$ and $\boldsymbol{p}_{in}$ given at the very beginning of the proof, we note that $\boldsymbol{p}_n = K_N^n \boldsymbol{z}_n$, where $K_N^n = V_n U_n^{-1}$, and $\boldsymbol{p}_{in} = K_N^{in} \boldsymbol{z}_{in}$, where $K_N^{in} = V_{in} U_{in}^{-1}$. Plugging these formulas into (4.6) and using (4.4), one easily gets

$$\frac{1}{\mu_N} \sigma_L(\{\hat{\boldsymbol{u}}_n\}, \boldsymbol{\ell}, T) = \boldsymbol{\ell} \cdot K_N(L) \boldsymbol{\ell} \,. \tag{4.7}$$

To prove (4.5), we recall that, by Corollary 3.7, $\sigma(\boldsymbol{u}, T, \boldsymbol{\ell}) = \mu_N K_N(T) \boldsymbol{\ell} \cdot \boldsymbol{\ell}$, provided $K_N = V U^{-1}$, where the matrix-valued functions V, U solve (3.25). We stress that the assumption $M_{jk} \equiv 0$ is precisely the necessary condition that (4.4) be a symplectic s-stage RK method (see, for instance, [10, p. 192]) for (3.25). In addition, we have $\|U(t_n) - U_n\|_2 = O(h^p)$ and $\|V(t_n) - V_n\|_2 = O(h^p)$, as the RK-method has order p by assumption. Now, we note that, although $U^{-1}(T)$ is well defined, it can be ill-conditioned numerically. To overcome this, we note that under the change of variables $U(t) := \widehat{U}(t)X$, $V(t) := \widehat{V}(t)X$, where $\widehat{U}$, $\widehat{V}$ solve (3.25), one would get that $\widehat{K_N}(t) = \widehat{V}(t)\widehat{U}^{-1}(t) = V(t)U^{-1}(t) = K_N(t)$. Therefore, we are free to reinitialize U_n, V_n at each time-step t_n, that is, we can compute K_N^{n+1} as $K_N^{n+1} = V_{n+1} U_{n+1}^{-1}$, where V_{n+1}, U_{n+1} are obtained through (4.4) with $V_n = K_N^n$ and $U_n = \mathcal{I}$. Computed in this way, U_L is well-conditioned, as it is close to the identity matrix $\mathcal{I}$ and so $\|U^{-1}(T) - U_L^{-1}\|_2 = O(h^p)$ implying that $\|K_N(T) - K_N^L\|_2 = O(h^p)$ for $K_N(T) = V(T)U^{-1}(T)$ and $K_N^n = V_n U_n^{-1}$. This and (4.7) proves the first equality in (4.5). Let us prove the second equality in (4.5). To this end, we recall that the minimax estimate $\hat{\boldsymbol{u}}(\boldsymbol{y}) = \boldsymbol{\ell} \cdot \hat{\boldsymbol{a}}_N(T)$ by Corollary 3.7, where $\hat{\boldsymbol{a}}_N$ solves (3.26). On the other hand, we note that (4.3) is an s-stage symplectic RK method for (3.26), and so one has at least $\|\hat{\boldsymbol{a}}_N(T) - \hat{\boldsymbol{a}}_N^L\|_{\mathbb{R}^N} = O(h)$. If $t \mapsto \boldsymbol{y}(t)$ is smooth, then the previous estimate can be improved. ∎

COROLLARY 4.2. *Assume that the coefficients a_{ij}, b_i are chosen so that the RK-method corresponds to a Gauss–Legendre method (see* [10, p. 34]*). Then the order of the method is $p = 2s$, and for $s = 1$ the discrete system* (4.3) *reads as the implicit midpoint rule:*

$$\begin{aligned} \hat{\boldsymbol{x}}_{1n} &= \hat{\boldsymbol{a}}_N^n - \frac{h}{2}(A_N(t_{n+\frac{1}{2}}) + \lambda_{N+1}^{\frac{1}{2}} S K_N^{1n} H_N'(t_{n+\frac{1}{2}}) H_N(t_{n+\frac{1}{2}}))\hat{\boldsymbol{x}}_{1n} \\ &\quad + \frac{h}{2} K_N^{1n} C_N'(t_{n+\frac{1}{2}}) R_N^{-1}(t_{n+\frac{1}{2}})(\boldsymbol{y}(t_{n+\frac{1}{2}}) - C_N(t_{n+\frac{1}{2}})\hat{\boldsymbol{x}}_{1n}) \,, \\ \hat{\boldsymbol{a}}_N^{n+1} &= 2\hat{\boldsymbol{x}}_{1n} - \hat{\boldsymbol{a}}_N^n \,, \quad \hat{\boldsymbol{a}}_N^0 = 0 \,, \quad t_{n+\frac{1}{2}} := t_n + \frac{h}{2} \,, \end{aligned} \tag{4.8}$$

where $K_N^{1n} = V_{1n} U_{1n}^{-1}$ and V_{1n}, U_{1n} solve (4.4) *with $s = 1$. If the coefficients a_{ij}, b_i are chosen so that the RK method corresponds to a diagonally implicit RK method of order p (see* [10, p. 147]*), then the s-stage method* (4.3) *may be represented as a composition of implicit midpoint steps* [10, p. 192].

Remark 3. Let us define the estimation error $\boldsymbol{\xi}(t) := \boldsymbol{a}_N^{true}(t) - \hat{\boldsymbol{a}}_N(t)$. Then, by differentiating $\boldsymbol{\xi}$ and using (3.1) and (3.26), it is easy to derive that

$$\frac{d\boldsymbol{\xi}}{dt} = -(A_N + K_N^{-1} D_N)\boldsymbol{\xi} + m(t) \,, \boldsymbol{\xi}(0) = \mathcal{P}_N I_0 \,,$$

where $D_N := \lambda_{N+1}^{\frac{1}{2}} S H_N' H_N + C_N' R_N^{-1} C_N$, $\boldsymbol{m} = \boldsymbol{e}^m + \mathcal{P}_N f - K_N C_N' R_N^{-1}(\boldsymbol{w} + \boldsymbol{e}) -$

$K_N\lambda_{N+1}^{\frac{1}{2}}SH'_N\boldsymbol{e}^o$. Now, by using (3.25), it is not hard to note that

$$\frac{dK_N^{-1}}{dt} = K_N^{-1}A_N + A'_NK_N^{-1} + D_N - K_N^{-1}Q_NK_N^{-1}\,, K_N^{-1}(0) = V(0)\,.$$

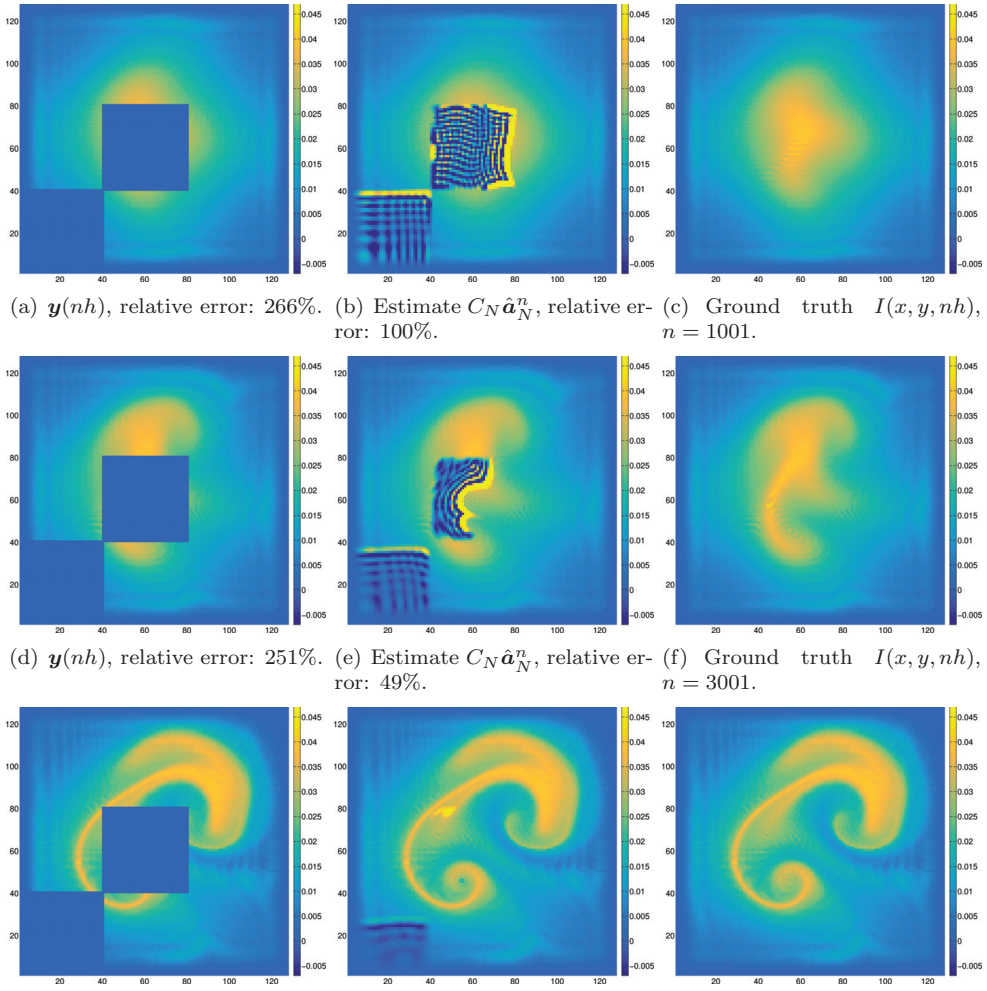
Then we have $\frac{d}{dt}K_N^{-1}\boldsymbol{\xi}\cdot\boldsymbol{\xi} = -(D_N + K_N^{-1}Q_NK_N^{-1})\boldsymbol{\xi}\cdot\boldsymbol{\xi} + 2K_N^{-1}\boldsymbol{\xi}\cdot m$, and so $K_N^{-1}\boldsymbol{\xi}\cdot\boldsymbol{\xi}$ decays along $t \mapsto \boldsymbol{\xi}(t)$ given that $2K_N^{-1}\boldsymbol{\xi}\cdot m$ is dominated by the quadratic term. Now, by using an argument of [8], it may be demonstrated that $(K_N)^{-1}\boldsymbol{\xi}_n\cdot\boldsymbol{\xi}_n$ decays.

5. Case study. As a proof of concept for the minimax projection method, in this section we compute an idealized experiment with specifications similar to real pollutant tracking problems. In particular, we assume that the observations of a discharged pollutant are available in the form of images in which observation data is either lacking or occluded by moving clouds, making it impossible to track and predict the pollutant from the image data only. Specifically, we will consider two test cases in which the observations differ. In Case I, we impose incomplete observations as well as a moving cloud profile over the domain. In Case II, we consider a situation in which observations have large error over part of the domain. We first describe Case I in detail, and then point out the differences with Case II.

5.1. Test Case I. The pollutant is discharged in the center of the domain $\Omega = (0, 2\pi)^2$. The initial concentration $I_0(x, y)$ is a radial Gaussian profile centered at (π, π) with standard deviation 2. The pollutant concentration $I(x, y, t)$ evolves according to the linear transport equation (2.1) with $\varepsilon = 0$. The fluid flow $\boldsymbol{v} = (u(x, y, t), v(x, y, t))'$ is computed by solving the 2D incompressible Euler equation in vorticity-stream function form as suggested in [11] with homogeneous Dirichlet boundary conditions for vorticity and stream functions. The initial vorticity field is obtained from the MATLAB `peaks` function. The vorticity field is then approximated by using a Fourier pseudospectral discretization on a uniform 128 x 128 grid, denoted Γ, with fourth-order explicit RK time-stepping. For each time-step, we project the vorticity field onto a span of eigenfunctions of the Laplacian, $\{\varphi_{ks} = \sin(\frac{kx}{2})\sin(\frac{sy}{2})\}$, that allows us to find the exact stream function by solving the Poisson equation.

The above approach yields a semianalytical representation of $\boldsymbol{v}(\boldsymbol{x}, nh)$ which is inserted in (2.1). The latter is then projected onto $\mathrm{span}\{\varphi_{ks}\}_{k,s=1}^{N^{\frac{1}{2}}}$ to compute the stiffness matrix $A_N(t)$. Model error f was represented as a linear combination of φ_{ks}, with random coefficients uniformly distributed in $(0, 1)$. Finally, the resulting nonstationary, nonhomogeneous linear system for the projection coefficients $\boldsymbol{a}_N^{true}$ was integrated in time using the implicit midpoint rule to obtain semianalytical representation for $I(x, y, t) = \sum_{k,s=1}^{N^{\frac{1}{2}}} a_{ks}^{\mathrm{true}}(t)\varphi_{ks}$. We used 55 basis functions in each direction (x, y) so that $N = 55^2$. Snapshots of $I(x, y, hn)$ are displayed in Figure 1(c)–1(i) for the case of $L = 8000$, $h = \frac{T}{L} \approx 0.0002$, and $T \approx 1.36$. Roughly speaking, the flow $\boldsymbol{v}$ is represented by two vortices which move clockwise inside the domain Ω and transform the initial concentration I_0 into a mushroom-like shape, as shown in the figures.

To generate observations $\boldsymbol{y}(hn)$, we projected the continuous in space snapshots $I(x, y, hn)$ onto the grid Γ, so that $\boldsymbol{y}(nh)$ is an $M := 128^2$-vector: $\boldsymbol{y}(nh) = I(x_i, y_j, nh)$ for x_i, y_j in Γ. Accordingly, we set $g_{ij} := \varepsilon^{-2}B(\frac{x-x_i}{\varepsilon})B(\frac{y-y_j}{\varepsilon})$, where B denotes the quadratic B-spline concentrated at 0, so that $\langle g_{ij}, I(\cdot, \cdot, nh)\rangle_{L^2(\Omega)} \approx I(x_i, y_j, nh)$. As a result, the $(k - 1 + s)$th column of C_N is composed of the values taken by φ_{ks} over the grid Γ. We define a region $\mathcal{O}_1 = \{(x, y)\,|\,x, y > \frac{5}{4}\}$ in the upper-right part of the grid, where observations are lacking, so C_N does not contain rows corresponding to that part. We also introduced nonstationary observational



(a) $\boldsymbol{y}(nh)$, relative error: 266%. (b) Estimate $C_N\hat{\boldsymbol{a}}_N^n$, relative error: 100%. (c) Ground truth $I(x,y,nh)$, $n=1001$.

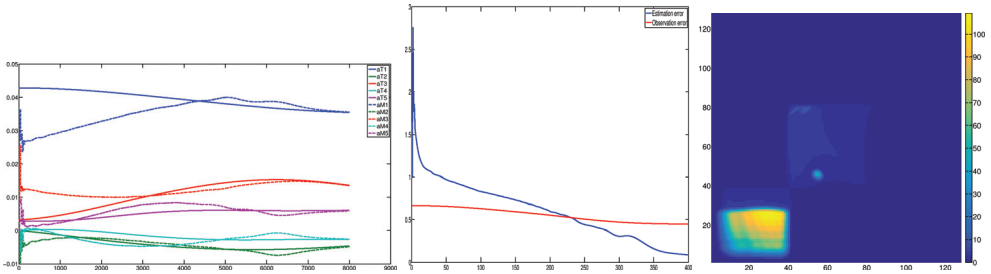
(d) $\boldsymbol{y}(nh)$, relative error: 251%. (e) Estimate $C_N\hat{\boldsymbol{a}}_N^n$, relative error: 49%. (f) Ground truth $I(x,y,nh)$, $n=3001$.

(g) $\boldsymbol{y}(nh)$, relative error: 224%. (h) Estimate $C_N\hat{\boldsymbol{a}}_N^n$, relative error: 20%. (i) Ground truth $I(x,y,nh)$, $n=8001$.

FIG. 1. *Test Case* I: *observed images* $\boldsymbol{y}(nh)$*, minimax estimates* $C_N\hat{\boldsymbol{a}}_N^n$*, and ground truth* $I(x,y,nh)$.

noise η in a form mimicking slowly translating clouds. The clouds are defined with respect to a periodic function composed of two Fourier modes, where the occluded regions are enclosed by a chosen level set. The clouds advect slowly over the domain with uniform wind vector $(1,1)$. In all, 400 observations were extracted (1 image per 20 time-steps) and occluded. The observed images $\boldsymbol{y}(nh)$ together with the relative norm of the resulting observation error, that is, $\frac{\|y(nh)-I(x_i,y_j,nh)\|_{\mathbb{R}^M}}{\|I(x_i,y_j,nh)\|_{\mathbb{R}^M}}$, are shown in Figure 1(a)–1(g). The weighting matrix R is set up so that the occluded regions have variance 10,000 and the rest of the observed image has variance 0.01. We assume that the "pixels" are uncorrelated according to the hyperbolic nature of (2.1) and so R is diagonal.

For the minimax projection method, we followed the procedure given in Remark 2. Namely, we checked that $\mu_N \approx 1$ and $\|H_N' H_N\|, \|(\mathcal{I}-\mathcal{P}_N^\dagger \mathcal{P}_N)g_{ij}\|_{L^2(\Omega)} \approx 0$ for $N=55^2$. We also ensured that (2.1) was well-resolved numerically, as can be assessed



(a) Estimates aM1 . . . aM5 vs. true coefficients aT1 . . . aT5. (b) Relative estimation and observation errors for images; (c) uncertainty map: minimax error in the image space.

FIG. 2. *Test Case* I: *convergence measures.*

visually looking at Figure 1(c)–1(i). For the state equation, we chose the diffusive version of the transport equation, (2.1) with $\varepsilon = 0.01$, which is equivalent to adding $\varepsilon\,\mathrm{diag}(\lambda_1 \dots \lambda_N)$ to the stiffness matrix $-A_N$ computed for (2.1), as described above. The latter introduces nonadditive model error which is taken into account together with additive model error f by setting $Q(\boldsymbol{x}, t) \equiv 1$. We also used $Q_0(x) \equiv 0.01\mathcal{I}$, reflecting the fact that we do not have any information about the initial condition (the place and amount of the discharge). The discrete minimax estimate $\hat{\boldsymbol{a}}_N^n$ was implemented by using the implicit midpoint rule (4.8), and the discrete gain $K_N(n+1)$ was computed using (4.4) with $s = 1$ and reinitialization $U_n = \mathcal{I}$, $V_n = K_N^n$ discussed in the proof of Proposition 4.1. As was suggested in Remark 2, we dropped all the terms involving $\lambda_{N+1}^{-\frac{1}{2}} S^{-1}$ and $\lambda_{N+1}^{-1} V^{-1}$.

The discrete filter $\hat{\boldsymbol{a}}_N^n$ starts from zero and $K_N(0) = 100\mathcal{I}$. Hence, the relative error in the initial condition is 100%. Observations $\boldsymbol{y}(nh)$ are assimilated at timesteps nh, $n = 21, 41, \dots, 8001$. For other n, we set $C_N(n) = 0$, which corresponds to the case with no innovation term. In the latter case $\hat{\boldsymbol{a}}_N^n$ evolves according to (4.8) with zero innovation term (2nd line in (4.8)). Since the observations are discrete in time and observation noise together with model error are nonstationary, the filter $\hat{\boldsymbol{a}}_N^n$ converges to the "true" projection coefficients only at the end of the time window: Figure 2(a) compares estimates of the first five projection coefficients against the truth. The estimation results (in the "space of images") are shown on Figure 1(b), 1(e), 1(h), where we can see how the convergence in the "space of coefficients" corresponds to the convergence in the "image's space."

We note that after the transition phase, the estimate reconstructs the solution occluded by clouds and in the unobserved region: $\frac{\|I(x_i,y_j,T) - C_N \hat{\boldsymbol{a}}_N^L\|_{\mathbb{R}^M}}{\|I(x_i,y_j,T)\|_{\mathbb{R}^M}} \leq 20\%$ (see Figure 1(g), 1(h), 1(i)). The reason for this is that the flow $\boldsymbol{v}$ is quite strong across the boundary of $\mathcal{O}_1$ and so the "trusted" observations from the adjacent regions flow into the unobserved region. The latter allows the filter to pick-up the right shape and the magnitude of the image in $\mathcal{O}_1$. We also observe that the model error is smoothed out, which explains the 20% relative error of the final estimate. Finally, the dynamics of the relative estimation error $\frac{\|I(x_i,y_j,T) - C_N \hat{\boldsymbol{a}}_N^L\|_{\mathbb{R}^M}}{\|I(x_i,y_j,T)\|_{\mathbb{R}^M}}$ is compared against the relative observation error in Figure 2(b): estimation error drops from 100% to 20% as opposed to the observation error, which stays above 200%.

5.2. Test Case II. The second test case simulates a scenario in which observations are unreliable in a part of the domain, e.g., due to an instrument failure. To

model this situation, the domain was partitioned into a 3×3 array, and two subdomains occluded: the centermost subdomain $\mathcal{O}_2 = \{(x,y) \,|\, \frac{5}{8}\pi \le x,y < \frac{5}{4}\pi\}$ and the lower left subdomain $\mathcal{O}_3 = \{(x,y) \,|\, x,y < \frac{5}{8}\pi\}$. This test case is challenging because a dynamically interesting part of the solution is obscured for much of the simulation.

For Case II, the true solution was computed at higher resolution, using a 75-mode truncation in each direction, i.e., $N = 75^2$. Furthermore, the model was assumed perfect: $f \equiv 0$. Imperfect observations of the square regions $\mathcal{O}_2$ and $\mathcal{O}_3$ were obtained from the discrete images $\boldsymbol{y}(nh)$ by setting $I(x_i, y_j, nh) = 0$ for $(x_i, y_j) \in \mathcal{O}_{2,3} \subset \Gamma$. The observed images $\boldsymbol{y}(nh)$ together with the relative norm of the resulting observation error are shown in Figure 3(a)–3(g). Observation uncertainty was again defined by diagonal R with occlusion patches having variance 10,000 compared to variance 0.01 elsewhere. We also used $Q_0(x) \equiv 0.01\mathcal{I}$ reflecting the fact that we do not have any information about the initial condition (the place and amount of the discharge), and $Q(\boldsymbol{x}, t) \equiv 100$, to indicate high confidence in our PDE model (2.1).

The estimation results (in the space of images) are shown in Figure 3(b), 3(e), 3(h). We note that again after the transition phase the estimate perfectly reconstructs the central occluded region $\mathcal{O}_2$, i.e., $\frac{\|I(x_i,y_j,T) - C_N \hat{\boldsymbol{a}}_N^L\|_{\mathbb{R}^M}}{\|I(x_i,y_j,T)\|_{\mathbb{R}^M}} \le 0.08$ (see Figure 3(g), 3(h), 3(i)) thanks to the very strong flow $\boldsymbol{v}$ over $\mathcal{O}_2$. In contrast, $\boldsymbol{v}$ is not strong in the lower-left region $\mathcal{O}_3$ and so the reconstruction is imperfect. This intuitive description is in full agreement with Figure 4(c), where the uncertainty map (minimax errors in the space of images) and the corresponding occlusion pattern are shown: as we can see, the uncertainty is quite high in $\mathcal{O}_3$ as opposed to $\mathcal{O}_2$. Finally, the relative estimation error drops from 100% to 8%, as opposed to the relative observation error, which stays above 45% (see Figure 4(b)).

6. Conclusion. In this paper we solve the state estimation problem for linear parabolic PDEs using a "discretize and optimize" strategy. That is, to project PDE (2.1) and its solution I onto a finite dimensional space, to bound the truncation error, and then derive the DAE for the projection coefficients. Using the minimax approach, we derive the state estimate for the DAE in the form of the linear filter (3.26), which depends on the number of the basis functions N and the norm of ΔI through the terms involving $\lambda_{N+1}^{-\frac{1}{2}} S^{-1}$ and $\lambda_{N+1}^{-1} V^{-1}$. Consequently, for large enough N, these terms have little or no impact and the constructed estimate converges to the infinite dimensional state estimator. We conclude that the "discretize and optimize" strategy adopted in the paper is equivalent to "optimize and discretize" in the limit $N \to \infty$.

Appendix A. Proofs of lemmas. In this appendix we provide the proofs of Lemmas 3.2, 3.3, and 3.4.

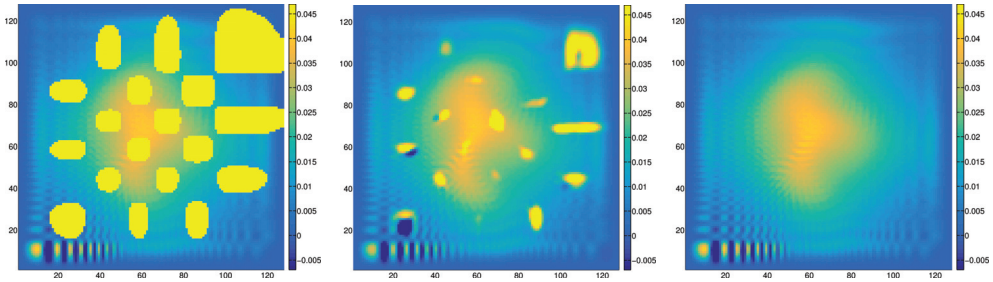
A.1. Proof of Lemma 3.2.

Proof. The first part of the claim follows from the standard results on second-order parabolic equations [7, p. 374]. To show that $\Delta I(\cdot, t) \in L^2(\Omega)$ for almost all $t \in (0,T)$, we employ the following assertion. For any $g \in L^2(0,T,L^2(\Omega))$, there exists the unique $I \in B(T) := L^2(0,T,H^2(\Omega) \cap H_0^1(\Omega))$ such that

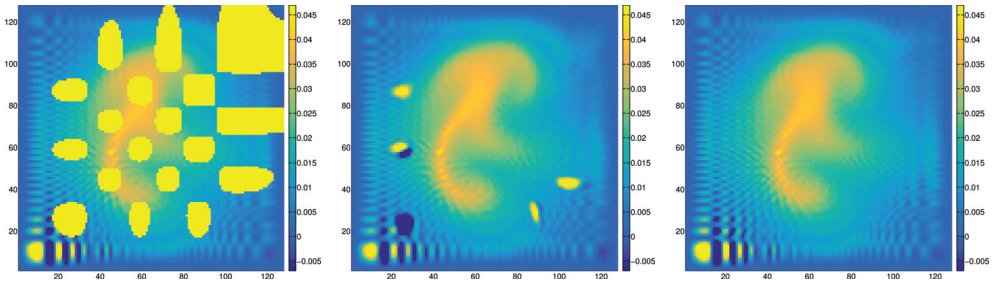
$$\partial_t I - \varepsilon \Delta I = g\,, I(\boldsymbol{x}, 0) = 0\,, \tag{A.1}$$

$$\|I\|^2_{L^\infty(0,T,H_0^1(\Omega))} \le (2\varepsilon)^{-1} \|g\|^2_{L^2(0,T,L^2(\Omega))}\,. \tag{A.2}$$

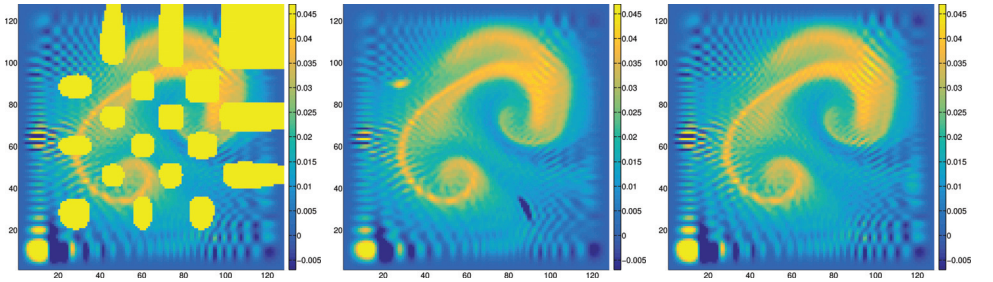
The existence and uniqueness of $I \in B(T)$ solving (A.1) was proved in [21]. The estimate (A.2) can be verified projecting (A.1) on the span of $\{\psi_k := \lambda_k^{-\frac{1}{2}} \varphi_k\}$, which



(a) $\boldsymbol{y}(nh)$, relative error: 66%. (b) Estimate $C_N\hat{\boldsymbol{a}}_N^n$, relative error: 90%. (c) Ground truth $I(x,y,nh)$, $n = 1001$.

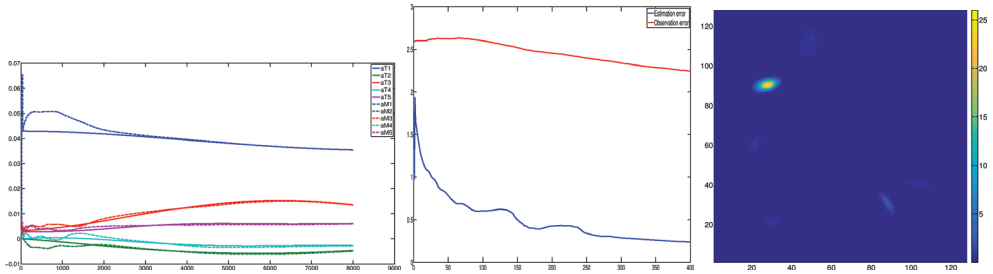


(d) $\boldsymbol{y}(nh)$, relative error: 59%. (e) Estimate $C_N\hat{\boldsymbol{a}}_N^n$, relative error: 67%. (f) Ground truth $I(x,y,nh)$, $n = 3001$.



(g) $\boldsymbol{y}(nh)$, relative error: 44%. (h) Estimate $C_N\hat{\boldsymbol{a}}_N^n$, relative error: 8%. (i) Ground truth $I(x,y,nh)$, $n = 8001$.

FIG. 3. *Test Case* II: *observed images* $\boldsymbol{y}(nh)$, *minimax estimates* $C_N\hat{\boldsymbol{a}}_N^n$, *and ground truth* $I(x,y,nh)$.



(a) Estimates aM1 . . . aM5 vs. true coefficients aT1 . . . aT5. (b) Relative estimation and observation errors for images; (c) uncertainty map: minimax error in the image space.

FIG. 4. *Test Case* II: *convergence measures.*

form an orthonormal basis in $H_0^1(\Omega)$ with respect to the inner product $\langle u, v\rangle_1 := \langle \nabla u, \nabla v\rangle_{L^2(\Omega)}$, and estimating the projection coefficients by applying the Cauchy–Schwatrz–Bunyakovsky inequality. Let us now prove that $\Delta I(\cdot, t) \in L^2(\Omega)$ for almost all $t \in (0, T)$. Indeed, we introduce a linear operator $v \mapsto I$ assigning $v \in X(T) := L^\infty(0, T, H_0^1(\Omega))$ the solution $I \in B(T)$ of (A.1), which corresponds to $g = g(v) := f - \boldsymbol{v} \cdot \nabla v \in L^2(0, T, L^2(\Omega))$. Then, by applying the same argument as in [7, p. 425], we prove that $v \mapsto I$ has a fixed point $I^* \in X(T^*)$ for small enough $0 < T^* \le T$. Therefore, the parabolic equation $\partial_t I + AI = f$, $I(\boldsymbol{x}, 0) = 0$ has a unique solution $I \in B(T^*)$ and so $\Delta I(\cdot, t) \in L^2(\Omega)$ for almost all $t \in (0, T^*)$. In the case $I(\boldsymbol{x}, 0) = I_0 \ne 0$, we have that $AI_0 \in L^2(0, T^*, L^2(\Omega))$ and so the PDE $\partial_t I_1 + AI_1 = f - AI_0$, $I_1(\boldsymbol{x}, 0) = 0$ has the unique solution $I_1 \in B(T^*)$. But then $I := I_1 + I_0$ solves $\partial_t I + AI = f$, $I(\boldsymbol{x}, 0) = I_0$, and $\Delta I(\cdot, t) = \Delta I_1 + \Delta I_0 \in L^2(\Omega)$ for almost all $t \in (0, T^*)$. Now, to conclude the proof, we split up the original interval $(0, T)$ into subintervals $(0, T^*)$, $(T^*, 2T^*)$, and so forth and repeat the above argument to prove that $\Delta I(\cdot, t) \in L^2(\Omega)$ for almost all $t \in (0, T)$. ∎

A.2. Proof of Lemma 3.3.

Proof. Noting that (see, for instance, [7, p. 357]) $\{\lambda_k^{-\frac{1}{2}}\varphi_k\}$ form an orthonormal basis in $H_0^1(\Omega)$ with respect to the inner product $\langle u, v\rangle_1 := \langle \nabla u, \nabla v\rangle_{L^2(\Omega)}$, and recalling that $I^N = \mathcal{P}_N^\dagger \boldsymbol{a}_N^{true} = \sum_{i=1}^N a_i \varphi_i$, we derive

$$\begin{aligned} \boldsymbol{e}^o \cdot \boldsymbol{e}^o &= \|H_N \boldsymbol{a}_N^{true}\|^2_{\mathbb{R}^N} = \|(\mathcal{I} - \mathcal{P}_N^\dagger \mathcal{P}_N) A I^N(\cdot, t)\|^2_{L^2(\Omega)} \\ &= \sum_{k>N} \langle \varphi_k, A I^N(\cdot, t)\rangle^2_{L^2(\Omega)} = \sum_{k>N} \langle \varphi_k, \boldsymbol{v}(\cdot, t) \cdot \nabla I^N(\cdot, t)\rangle^2_{L^2(\Omega)} \\ &= \sum_{k>N} \lambda_k^{-2} \langle -\Delta \varphi_k, \boldsymbol{v}(\cdot, t) \cdot \nabla I^N(\cdot, t)\rangle^2_{L^2(\Omega)}. \end{aligned} \tag{A.3}$$

Now, we claim that

$$\langle -\Delta \varphi_k, \boldsymbol{v}(\cdot, t) \cdot \nabla I^N(\cdot, t)\rangle^2_{L^2(\Omega)} = \langle \varphi_k, \boldsymbol{v}(\cdot, t) \cdot \nabla I^N(\cdot, t)\rangle_1. \tag{A.4}$$

Indeed, it is sufficient to apply integration by parts (3.9) to the left-hand side of (A.4) and note that $\mathrm{tr}(\boldsymbol{v}(\cdot, t) \cdot \nabla I^N(\cdot, t)) = 0$. The latter can be shown, in turn, by approximating $M_i(\cdot, t) \in H_0^1(\Omega)$ with smooth functions $\psi_i^j \in C_c^\infty(\Omega)$ such that $\lim_{j\to\infty} \|\psi_i^j - M_i(\cdot, t)\|_{H^1(\Omega)} = 0$ and $\mathrm{tr}(\psi_i^j) = 0$ for almost all t (see [7]). Now, combining (A.4) with (A.3), we get

$$\begin{aligned} \boldsymbol{e}^o \cdot \boldsymbol{e}^o &= \lambda_{N+1}^{-1} \sum_{k>N} \frac{\lambda_{N+1}}{\lambda_k} \langle \lambda_k^{-\frac{1}{2}} \varphi_k, \boldsymbol{v}(\cdot, t) \cdot \nabla I^N(\cdot, t)\rangle_1^2 \\ &\le \lambda_{N+1}^{-1} \|\boldsymbol{v}(\cdot, t) \cdot \nabla I^N(\cdot, t)\|_1^2. \end{aligned} \tag{A.5}$$

Let us now estimate the last term in (A.5). Equation (3.10) implies

$$\|\nabla I^N(\cdot, t)\|^2_{L^2(\Omega)} \le \lambda_1^{-1} \|\Delta I^N(\cdot, t)\|^2_{L^2(\Omega)} \le \lambda_1^{-1} \|\Delta I(\cdot, t)\|^2_{L^2(\Omega)}. \tag{A.6}$$

Denote $\boldsymbol{v}^j := (\psi_1^j \dots \psi_n^j)'$, where ψ_i^j is a smooth function approximating M_i as suggested above. Then $\boldsymbol{v}^j \in C_c^\infty(\Omega)$, and consequently we can write

$$\|\boldsymbol{v}^j \cdot \nabla I^N(\cdot, t)\|_1^2 = \sum_{i=1}^n \int_\Omega \left(\sum_{k=1}^n \partial_{x_i} \psi_k^j \partial_{x_k} I^N(\boldsymbol{x}, t) + \psi_k^j \partial^2_{x_i x_k} I^N(\boldsymbol{x}, t) \right)^2 d\boldsymbol{x}.$$

Now, recalling that $\sum_{i,k=1}^{n}\int_\Omega(\partial^2_{x_ix_k}I^N(\boldsymbol{x},t))^2d\boldsymbol{x}\le\|\Delta I^N(\cdot,t)\|^2_{L^2(\Omega)}$ for a convex open bounded domain Ω (see [1, 13]), we bound $\|\boldsymbol{v}^j\cdot\nabla I^N(\cdot,t)\|_1^2$ by applying the Cauchy–Schwarz–Bunyakovsky inequality:

$$\|\boldsymbol{v}^j\cdot\nabla I^N(\cdot,t)\|_1^2\le 2\int_\Omega(\|J_{\boldsymbol{v}^j}(\boldsymbol{x},t)\|_2^2\|\nabla I^N(\boldsymbol{x},t)\|^2_{\mathbb{R}^n}+\|\boldsymbol{v}^j(\boldsymbol{x},t)\|^2_{\mathbb{R}^n}\|\Delta I^N(\boldsymbol{x},t)\|^2_{\mathbb{R}^N})d\boldsymbol{x}\,.$$

Now, recalling that $\boldsymbol{v}^j\to\boldsymbol{v}(\cdot,t)$ in $H^1(\Omega)$ and taking limits in the above inequality, we deduce that it holds true for $\boldsymbol{v}(\cdot,t)\in H_0^1(\Omega)$ (for almost all t). This latter observation, (A.6), and (A.3) prove (3.3). ∎

A.3. Proof of Lemma 3.4.

Proof. Let I solve (2.1). Then, $\varepsilon\Delta I=\partial_tI+\boldsymbol{v}\cdot\nabla I-f$ in Ω and so

$$\begin{aligned}\varepsilon^2\|\Delta I(\cdot,t)\|^2_{L^2(0,T,L^2(\Omega))}&=\int_0^T\langle\partial_tI+\boldsymbol{v}\cdot\nabla I-f,\varepsilon\Delta I\rangle_{L^2(\Omega)}\,dt\\&\le 3\int_0^T\|f\|^2_{L^2(\Omega)}+\|\rho_1(\cdot,t)\|_{L^\infty(\Omega)}\|\nabla I(\cdot,t)\|^2_{L^2(\Omega)}\\&\quad+\|\partial_tI\|^2_{L^2(\Omega)}\,dt\,.\end{aligned}\tag{A.7}$$

Applying the energy method [7, p. 372], we obtain

$$\begin{aligned}&\int_0^T\|\partial_tI^N\|^2_{L^2(\Omega)}\,dt+\varepsilon\|\nabla I^N(\cdot,T)\|^2_{L^2(\Omega)}\\&\quad\le\varepsilon\|\nabla I^N(\cdot,0)\|^2_{L^2(\Omega)}\\&\qquad+2\|f\|^2_{L^2(0,T,L^2(\Omega))}+2\int_0^T\|\rho_1(\cdot,t)\|_{L^\infty(\Omega)}\|\nabla I^N(\cdot,t)\|^2_{L^2(\Omega)}\,dt\,.\end{aligned}\tag{A.8}$$

By (A.6), we get $\|\nabla I^N(\cdot,0)\|^2_{L^2(\Omega)}\le\|\nabla I_0\|^2_{L^2(\Omega)}$ and $\|\nabla I^N(\cdot,t)\|_{L^2(\Omega)}\le\|\nabla I(\cdot,t)\|_{L^2(\Omega)}$. This and (A.8) imply that the sequence $\{\partial_tI^N\}_{N\in\mathbb{N}}$ is bounded in $L^2(0,T,L^2(\Omega))$, and so we can find a subsequence $\{\partial_tI^{N_k}\}$ weakly converging to ∂_tI in $L^2(0,T,L^2(\Omega))$. As the norm in $L^2(0,T,L^2(\Omega))$ is weakly lower-semicontinuous, we get

$$\begin{aligned}\int_0^T\|\partial_tI\|^2_{L^2(\Omega)}\,dt&\le\lim\int_0^T\|\partial_tI^{N_k}\|^2_{L^2(\Omega)}\,dt\le 2\|f\|^2_{L^2(0,T,L^2(\Omega))}\\&\quad+\varepsilon\|\nabla I_0\|^2_{L^2(\Omega)}+2\int_0^T\|\rho_1(\cdot,t)\|_{L^\infty(\Omega)}\|\nabla I(\cdot,t)\|^2_{L^2(\Omega)}\,dt\,.\end{aligned}\tag{A.9}$$

Since (A.8) holds for any $T>0$, it follows by the Gronwall inequality in the integral form that

$$\|\nabla I^N(\cdot,t)\|^2_{L^2(\Omega)}\le\left(\frac{2}{\varepsilon}\|f\|^2_{L^2(0,T,L^2(\Omega))}+\|\nabla I_0\|^2_{L^2(\Omega)}\right)\exp\left\{\int_0^t\frac{2}{\varepsilon}\|\rho_1(\cdot,t)\|_{L^\infty(\Omega)}ds\right\},$$

and by the weak convergence argument, we get the same estimate for $\nabla I(\cdot,t)$. Combining this latter estimate with (A.7) and (A.9) gives (3.4). ∎

REFERENCES

[1] V. Adolfsson, *L^2-integrability of second order derivatives for Poisson equation in nonsmooth domain*, Math. Scand., 70 (1992), pp. 140–160.

[2] J. Aubin, *Approximation of Elliptic Boundary-Value Problems*, Wiley, New York, 1972.

[3] J. Baumeister, W. Scondo, M. Demetriou, and I. Rosen, *On-line parameter estimation for infinite-dimensional dynamical systems*, SIAM J. Control Optim., 35 (1997), pp. 678–713.

[4] A. Bensoussan, *Filtrage Optimal des Systèmes Linéaires*, Dunod, Paris, 1971.

[5] F. L. Chernousko, *State Estimation for Dynamic Systems*, CRC, Boca Raton, FL, 1994.

[6] R. F. Curtain and A. J. Pritchard, *Infinite Dimensional Linear Systems Theory*, Springer, New York, 1978.

[7] L. Evans, *Partial Differential Equations*, 2nd ed., Grad. Stud. Math. 19, AMS, Providence, RI, 2010.

[8] J. Frank and S. Zhuk, *Symplectic Möbius integrators for LQ optimal control problems*, in Proceedings of the IEEE Conference on Decision and Control, 2014, pp. 6377–6382.

[9] P. Grisvard, *Elliptic Problems in Non-smooth Domains*, Pitman, London, 1985.

[10] E. Hairer, C. Lubich, and G. Wanner, *Geometric Numerical Integration*, 2nd ed., Springer, New York, 2006.

[11] I. Herlin, D. Bereziat, N. Mercier, and S. Zhuk, *Divergence-free motion estimation*, Computer Vision—ECCV 2012, Lecture Notes in Comput. Sci. 2012, Springer, 2012, pp. 15–27.

[12] J. Hesthaven, S. Gottlieb, and D. Gottlieb, *Spectral Methods for Time-Dependent Problems*, Cambridge University Press, Cambridge, 2007.

[13] J. Kadlec, *On the regularity of the solution of the Poisson problem in a domain with boundary locally similar to the boundary of a convex open set*, Czechoslovak Math. J., 14 (1964), pp. 386–393.

[14] A. Kurzhanski and I. Vályi, *Ellipsoidal Calculus for Estimation and Control*, Systems Control Found. Appl., Birkhäuser, Boston, MA, 1997.

[15] J. L. Lions, *Optimal Control of Systems Governed by Partial Differential Equations*, Springer, New York, 1971.

[16] K. Morton and E. Süli, *Evolution-Galerkin methods and their supraconvergence*, Numer. Math., 71 (1995), pp. 331–355.

[17] A. Nakonechny, *A minimax estimate for functionals of the solutions of operator equations*, Arch. Math. (Brno), 14 (1978), pp. 55–59.

[18] A. Pazy, *Semigroups of Linear Operators and Applications to Partial Differential Equations*, Springer, New York, 1992.

[19] T. Reid, *Riccati Differential Equations*, Academic Press, New York, 1972.

[20] V. Thomée, *Galerkin Finite Element Methods for Parabolic Problems*, Comput. Math., Springer, New York, 1997.

[21] I. Wood, *Maximal L_p-regularity for the Laplacian on Lipschitz domains*, Math. Z., 255 (2007), pp. 855–875.

[22] S. Zhuk, *Closedness and normal solvability of an operator generated by a degenerate linear differential equation with variable coefficients*, Nonlinear Oscil., 10 (2007), pp. 464–480.

[23] S. Zhuk, *Estimation of the states of a dynamical system described by linear equations with unknown parameters*, Ukrainian Math. J., 61 (2009), pp. 214–235.

[24] S. Zhuk, *Minimax state estimation for linear discrete-time differential-algebraic equations*, Automatica J. IFAC, 46 (2010), pp. 1785–1789.

[25] S. Zhuk, *Minimax state estimation for linear stationary differential-algebraic equations*, in Proceedings of the 16th IFAC Symposium on System Identification, 2012, pp. 143–148.

[26] S. Zhuk, *Kalman duality principle for a class of ill-posed minimax control problems with linear differential-algebraic constraints*, Appl. Math. Optim., 68 (2013), pp. 289–309.

[27] S. Zhuk, *Minimax projection method for linear evolution equations*, in Proceedings of the IEEE Conference on Decision and Control, 2013, pp. 2556–2561.